

The international journal of science / 23 April 2020

index
Cancer

nature

FIRE AND RAIN

How extreme rainfall might
have helped trigger the 2018
Kilauea eruption

Climate scenarios

Fresh approaches to
modelling the future
of planet Earth

Car trouble

The cost errors
stalling the switch
to green vehicles

Rate of change

Predicting how fast
biodiversity will alter
in a warming world

Vol. 580, No. 7834
nature.com

Freezing World Health Organization funding is dangerous

Researchers everywhere must challenge Donald Trump's undermining of the global health agency.

Amid the biggest global health crisis in a century, it is dangerous to hurt the one intergovernmental agency most able to guide the world out of it. But that is precisely what happened this week.

On the orders of President Donald Trump, the United States, one of the key founding members of – and the largest donor to – the World Health Organization (WHO), announced its intention to freeze its funding for the agency, pending a review of WHO actions during the coronavirus pandemic. The review could last for up to three months.

Trump's administration has been increasingly critical of the agency, which for months has guided the world in how to tackle the deadly coronavirus. That work – and the WHO's other life-saving interventions around the world – will be at risk if the agency loses its US funding, which amounted to nearly US\$900 million for 2018–19.

It isn't yet clear whether the White House can withhold this funding – especially the portion that has been approved by the US Congress – and if so, how much it can keep back. But even talk of doing so in the middle of a global health and economic crisis cannot be condemned strongly enough.

De-funding the WHO is especially dangerous for those low-income countries in which the agency's work is crucial to maintaining standards of public-health infrastructure, and also to tackling killer diseases. The WHO's epidemiologists, clinicians and logistics personnel are right now overseeing more than 35 emergency operations, including a measles outbreak in the Democratic Republic of the Congo and a cholera outbreak in Yemen.

On top of its emergency operations, the WHO handles ongoing efforts to treat tuberculosis and diabetes; eradicate polio; and study tropical diseases. This is all on an annual budget of roughly \$2.4 billion. Of this, the WHO's emergency-response budget is approximately \$280 million. By contrast, the agency that tackles public-health emergencies in the United States – the Centers for Disease Control and Prevention – has a total budget of around \$12.7 billion this year.

Finding the balance

A pandemic is always a big test for the WHO. In previous health emergencies, the agency has been criticized for acting too slowly, or – in the case of the 2009 H1N1 influenza

pandemic – overstating the risks. But leading public-health researchers and practitioners agree that, so far in the current crisis, the agency has offered leadership and acted according to the evidence it has received.

The WHO was notified of a cluster of pneumonia cases by China on 31 December, and it began an emergency-response process the following day. Its many actions since then include posting and updating guidance on how to diagnose COVID-19, vetting diagnostic tests and distributing them around the world. The agency's science division convened world experts to survey potential therapeutics. From this, it developed an adaptable clinical-trial protocol, known as SOLIDARITY, that has been launched globally.

More recently, the WHO has set up a supply-chain management system to try to ensure that low-income countries are not left without tests, medical equipment or protective gear for health workers – given the fierce competition for these limited resources.

The WHO declared a public health emergency of international concern, or PHEIC, on 30 January. That announcement is a trigger for the agency's member governments to follow its recommendations. These include establishing a comprehensive programme of testing, quarantining people suspected to be infected, and tracing their contacts.

Some countries acted quickly, including Germany, Singapore and South Korea. But the United States is among those that has not followed these particular recommendations. Even now, it does not have a national infrastructure for testing for the virus, nor for tracing the contacts of those infected with it.

In early March, WHO director-general Tedros Adhanom Ghebreyesus pleaded with the world when he said: "You can't fight a virus if you don't know where it is. That means robust surveillance to find, isolate, test and treat every case, to break the chains of transmission."

But the Trump administration chose not to follow the WHO's advice. Instead, influential lawmakers have been calling for an investigation into the WHO's actions, claiming that the agency was too slow to sound the alarm and too deferential to the Chinese government. At the same time, they are implicating the WHO in wider questions being directed at China's government. These include that China could have acted more quickly to lock down in the days after the first outbreak, and that public officials withheld important information. Such questions must be asked of China, but they are not for the WHO – which acts at the behest of governments – to answer. And they are not reasons to de-fund the agency.

It is, of course, crucial that lessons are learnt from all stages of this pandemic. Once it is over, there will be many national and international investigations and inquiries – including the WHO's own – and these will uncover what went right, what went wrong and what could have been done better. It is always tough to operate in a pandemic, and tougher still when essential cooperation between governments is at a low ebb. Such inquiries will be an opportunity to improve and to grow. They are not a reason to undermine or attack.

This pandemic needs the world to follow a coordinated

 **The agency has offered leadership and acted according to the evidence it has received."**

plan, covering decisions including how and when lockdowns are to be relaxed. It is extraordinary that more than three months into the outbreak, such a plan is nowhere to be seen. On 19 April, the health ministers of the G20 group of nations met virtually. They must put such a plan in place. The best hope of achieving that is for all nations to work with the WHO and other international agencies.

It is right that researchers, funders and governments have been protesting against Trump's decision, and they must continue to do so in the strongest terms. Those in the United States must also lobby their lawmakers at every level. The president and his administration must not withhold funding from the WHO. Doing so will place more lives at risk and ensure that the world takes longer to emerge from this crisis.

Nearly 70 years ago, the United States was instrumental in helping to establish the WHO. Nations realized that they needed such an agency in part because they couldn't tackle pandemics by acting alone. It is a sad indictment of the state of our world that the agency is now having to fight for its future while doing the job it was created to do.

We need to support the WHO so it is at its strongest, not undermine it at such a crucial hour.

Climate action and poverty alleviation go hand-in-hand

The world urgently needs a post-pandemic consensus on tackling climate change.

For the first time since its inception 50 years ago, this year's Earth Day, on 22 April, will coincide with the fleeting prospect of a lower carbon footprint, as the fastest economic slowdown the world has ever seen has grounded transport and closed workplaces.

The 'new normal' – as some are calling it – also comes at huge social and economic cost. As *Nature* went to press, the SARS-CoV-2 coronavirus had taken more than 170,000 lives, a number that will continue to rise. And the pandemic has also precipitated an unprecedented economic shock. Worldwide, tens of millions have been made unemployed. For now, governments are rightly focusing on spending trillions of dollars to keep health-care systems functioning, to pay for rising welfare costs and to support companies to prevent more workers losing their jobs.

But, at the same time, many carbon-intensive industries in coal, oil and gas are queuing up for bailouts. Governments need to resist. Before the pandemic, momentum was building towards decarbonization – for example, through commitments from governments on net-zero

“
The pandemic has taught the world a sharp lesson in what happens when there is a swift economic shock.”

emissions and through green new deals. This work must not be undone.

But a greener post-pandemic future cannot come at the expense of livelihoods – particularly those of the lowest paid and those in developing countries. The United Nations is forecasting that a drop in demand from high-income nations means that low- and middle-income countries will lose hundreds of billions of dollars in export earnings in 2020. Without urgent research and action, many of these countries are looking at vast numbers of their citizens staying out of work.

Polluter pays

Fortunately, there's one action that could contribute to easing some of the coming hardships and, at the same time, ensure that development continues on a sustainable path. After the 1992 Rio Earth Summit, developed nations pledged to help developing nations with research and development and with green financing. This wasn't aid so much as an application of the 'polluter pays' principle. Many of the richer countries had recognized that their actions had caused climate change. And they agreed that they had a responsibility to fund less-developed countries, both to help those nations become more resilient to the effects of global warming, and so that those countries could continue to develop, albeit in greener ways.

A decade ago, developed countries pledged to channel US\$100 billion annually to developing nations in climate finance by 2020. But – as we reported in September (*Nature* 573, 328–331; 2019) – only \$71 billion reached its destination in 2017, and this was mostly in loans, not grants. In the context of today's bailouts, these are not onerous sums. Worldwide, some \$2.4 trillion a year will be needed for the next 15 years just to transform energy systems to keep global temperatures from exceeding 1.5 °C above pre-industrial levels. As the economic crisis deepens, more loans are being offered by multilateral lending agencies. But loans are no substitute for the failure to keep past promises.

It's unfortunate that the next Conference of the Parties to the UN's climate convention – due to take place in Glasgow, UK, in November – has had to be postponed, because this is where developed nations would have been reminded of their obligations. However, in the spirit of current work-pattern adjustments, this meeting – or at least preparations for it – could still take place virtually. The coming economic stimulus packages must include finance for greener development. And long-promised funding for developing countries must also be made good.

The pandemic has taught the world a sharp lesson in what happens when there is a swift economic shock. A similar shock could lie ahead – as economists have long warned – if action is not taken to curb climate change. The International Monetary Fund is projecting that growth in most countries is likely to bounce back in 2021 if lockdowns do not persist. But the world might not be so resilient should such a shock result from extreme climate events, or rising sea levels.

That is why greener forms of growth must remain a priority. But development must be equitable, too.

World view

To mark the 50th Earth Day, take collective action



By Emma Marris

Reducing your own carbon footprint is not as powerful as calling governments and companies to account.

Everybody can and should do something to protect our shared home, Earth. But it pays to be strategic when deciding what action to take. In recent years, I have joined the board of a local climate non-profit group, marched, written local and national editorials, and even been hauled to jail for occupying the Oregon state capitol to protest against a pipeline project. And I've never felt more sure of myself.

For at least the past decade, environmental groups have been suggesting we reduce individual emissions by altering our behaviour: choosing public transport, eating less meat, buying more efficient light bulbs. In fact, it was the oil giant BP that popularized the 'personal carbon footprint' in 2005. This focus has kept individuals working on their own impact while letting governments, and corporations that profit from climate change, off the hook.

This year, 22 April marked the 50th annual Earth Day, a day of protest in favour of environmental protection. It is past time to shift our focus to policies that can get at the root cause of the problem: the extraction and burning of fossil fuels. The transition to zero-emissions energy needs to happen quickly – but also equitably, without making energy unobtainable for poor and marginalized people. By acting together to demand this, we can all have an impact much larger than whatever reductions we can make as individuals.

The average person in an industrialized country is responsible for around 10 tonnes of carbon emissions per year, so that's a rough limit to what anyone can accomplish by addressing only their own footprint. If someone were to join a campaign to close down a coal-powered plant, and it prevailed, that could help to eliminate 10 million tonnes of carbon a year – much more than any individual could prevent over a whole lifetime. And helping to pass laws that shut down all coal plants in a single country could multiply that impact.

Scientists are well positioned to contribute to such campaigns. They have skills that can be valuable in this fight, even if their area of expertise has nothing to do with atmospheric chemistry, soil science, electrical engineering, conservation planning or any of the hundreds of other specialisms that directly intersect with the problem.

Scientific habits of mind can help anyone to identify where they have the most leverage. Do you join 10,000 people to push for a national policy change, 1,000 people to demand your employer divest from fossil fuels, or 10 people to push for a new bike lane by your house? Any choice will be good, and success at any level will do much

It pays to be strategic when deciding what action to take."

more to stop climate change than a lifetime of green living.

Everyone who works for an employer, is part of a community, or is a member of a field has several possible spheres of influence. They can push their employer to divest from fossil fuels; support students who are demanding changes across a university system; lobby a government agency to include climate impacts in decision-making; or even go on strike if their employer continues to pursue profits that are tangled up with torching the climate. They can go to council meetings and recommend that their community sets emissions targets or invests in community-run renewable energy or public transport. They can push for their professional organizations to make commitments and public statements. They can take to the streets and protest; they can run for office; they can volunteer to take the muffins to the next meeting of the local climate-activism group.

Climate change is not the only environmental problem facing Earth – there are other major threats to the diversity of life. Working to institute policies that protect complex ecosystems from clearance and development can also have high rewards. There might be species on the edge of extinction in your area that you can help by attending community meetings, writing comments during government planning processes or lobbying your local representatives.

Equally, fighting for the rights of people of colour, Indigenous people, people from sexual and gender minorities, poor people and members of other marginalized groups is a powerful way of engaging in the fight for climate justice. When everyone really has equal power, policies that allow historically favoured groups to profit from the ongoing and deepening misery of others will be changed.

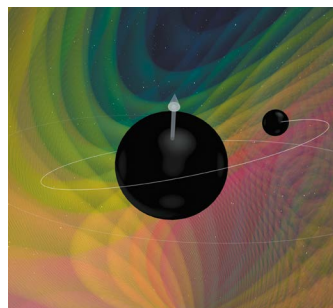
There are as many ways to engage with collective action as there are people. But the key word is collective. By all means, green up your personal life. It can help, and often comes with other benefits, as well. But we should prioritize changing the systems we all live in that make it difficult and expensive to make green choices.

You might worry that advocating on behalf of the climate puts your reputation as a dispassionate, rational scientist at risk. I sympathize. As a freelance writer, I worried that getting involved in climate activism would mean I could no longer claim to be an unbiased journalist. In the end, I decided that sharing my fears about the future, my love for Earth and its life, and my opinions about the changes we need to make to protect our home did not conflict with my core commitment to seek out and report the truth.

Going public with your opinions need not conflict with the scientific search for truth. Nor do you have to dedicate your life to collective action for the environment. Even a couple of hours every other week can achieve great things. Individually, we are puny; together, we can change the world.

Emma Marris
writes about the environment from Oregon.
e-mail: e.marris@gmail.com

News in brief



RARE BLACK-HOLE COLLISION OPENS NEW WINDOW ON UNIVERSE

Gravitational-wave astronomers have for the first time detected a collision between two black holes of substantially different masses – opening up a new vista on astrophysics and on the physics of gravity. The event offers the first unmistakable evidence from these faint space-time ripples that one black hole was spinning before the merger, giving astronomers rare insight into this key property.

“It’s an exceptional event,” said Maya Fishbach, an astrophysicist at the University of Chicago in Illinois, who unveiled the signal at a virtual meeting of the American Physical Society on 18 April. The US-based Laser Interferometer Gravitational-Wave Observatory (LIGO) and the Virgo observatory near Pisa, Italy, detected the collision last year.

The observation network has seen mergers between black holes with roughly equal masses. Physicists had eagerly awaited events with black holes of uneven mass because they provide more precise ways of testing the general theory of relativity. In the latest merger, one black hole was around 8 solar masses, and the other about 31. This imbalance made the larger black hole distort the space around it, so the other’s trajectory deviated from a perfect spiral. This could be seen in the resulting wave signal.

CORONAVIRUS DISRUPTS RESEARCH FUNDING

Although science is crucial to the fight against COVID-19, researchers unable to work are becoming increasingly concerned about how the coronavirus pandemic will affect their funding.

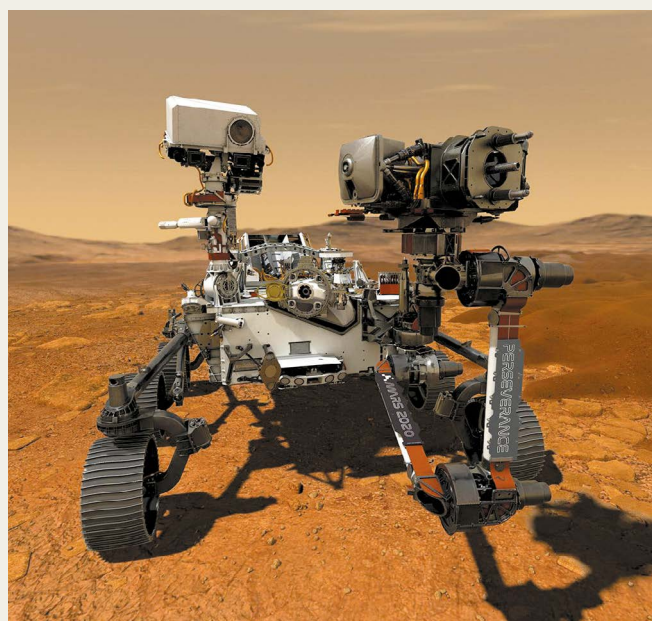
Disruptions such as the closure of labs and universities mean that researchers face challenges in completing projects and in paying lab members when grants run out.

“If this situation lasts for more than two to three months it will be impossible to finish the projects on time,” says Juan Astorga-Wells, a biochemist at the Karolinska Institute in Stockholm who is involved in two projects supported by the European Union’s Horizon 2020 research programme.

Many of the world’s major research funders have put in place policies to support grant-holders. Horizon 2020’s guidance says that researchers can ask to extend projects by up to six months, and reallocate funds to meet the costs of working remotely, or to help pay the salaries of researchers who are unable to continue with experiments. Projects can also be reoriented towards research on COVID-19 or coronaviruses.

US government funders including the National Institutes of Health also offer grant extensions, and will accommodate late applications for funding.

Funders have also put out rapid calls for proposals for coronavirus research. UK Research and Innovation has allocated £20 million (US\$25 million) to the topic.



Revealed: how a spacecraft will bring Mars rocks to Earth

NASA and the European Space Agency (ESA) have unveiled details of their upcoming mission to retrieve rock samples from Mars.

The first stage will start in July, when NASA launches its Perseverance rover (pictured) to roll around on the Martian surface and collect dust and rock. Perseverance will land next February in the red planet’s Jezero crater. As it drives around for many kilometres, the rover will drill or scoop up material to fill around 30 geological sampling tubes.

Until now, it hadn’t been clear how those tubes might get back to Earth. But after four years of designing and plotting, NASA and ESA have finalized a plan that involves sending a pair of spacecraft to Mars in 2026.

The first spacecraft would land in Jezero crater. There, a small rover would make its way to Perseverance, pick up the filled sampling tubes and transfer them to a “Mars ascent vehicle” – essentially a small rocket with a container to carry the samples. The Mars ascent vehicle would blast off and place the container into Martian orbit. The second spacecraft would then manoeuvre itself next to the sample container, pick it up and fly it back to Earth, probably landing in a military training ground in Utah.

If the plan works, scientists will finally get their hands on the Martian rocks in 2031.

News in focus



XINHUA NEWS AGENCY/SHUTTERSTOCK

Research from China is crucial to understanding the COVID-19 pandemic.

CHINA IS TIGHTENING ITS GRIP ON CORONAVIRUS RESEARCH FINDINGS

Some scientists welcome government vetting because it could stop publication of poor-quality COVID-19 papers – others fear it is an attempt to control information.

By Andrew Silver & David Cyranoski

China's government has started asserting tight control over COVID-19 research findings. Over the past two months, it seems to have quietly introduced policies that require scientists to get approval to publish – or publicize – their results, according to documents seen by *Nature* and some researchers.

This fits with media reports that at least two Chinese universities have posted notices online stating that research on the virus's origins needs approval from a university committee and the Ministry of Science and Technology (MOST) or Ministry of Education (MOE) before being submitted for publication.

Scientists in China say the changes are probably a response to poor-quality studies on the virus – and several welcome them.

But some academics have suggested that the vetting process could delay publication of important insights that could help to rein in the pandemic, and that the policies are part of China's attempt to control information about the start of the outbreak.

Last month, China's foreign-ministry spokesperson, Zhao Lijian, made sensational claims that the virus might have come to the country from the United States, prompting concerns that the Chinese government's statements were not always guided by science. Although the exact origin of the virus is unknown, researchers think it probably came

from bats and then spread to a carrier animal before infecting the first people somewhere in central China late last year.

Awareness of the new rules is mixed among researchers in China. The ministries seem not to have posted notices about the policies on their websites, and they have not yet responded to *Nature's* attempts to confirm that they have released the documents.

Government oversight of COVID-19 research seems to have started with a directive to universities. A document that seems to be from the MOE, and is dated 10 March, orders institutions to get approval from the ministry and the Joint Prevention and Control Mechanism, run by the powerful State Council, before publicly announcing results on the origin of the

SARS-CoV-2 virus, its transmission routes or treatments or vaccines. The document states that universities need to consider “the questions society is concerned about” when publicizing research on the virus. (*Nature* was sent the document, which is stamped by the MOE and includes the name of an agency official, by a researcher who did not want to comment.)

The education ministry seems to have issued another order after a meeting of the Joint Prevention and Control Mechanism on 25 March, according to a second notice that also seems to come from the MOE and has been posted on Pincong, a Chinese-language forum. This notice, dated 7 April, states that studies on the virus’s source must be approved by a university academic committee and the education ministry’s science and technology department before being published in a journal or posted on a preprint server or blog. Academic committees must evaluate all other COVID-19 papers for “academic value and timing”. The notice also warns that studies must not exaggerate the efficacy of vaccines or treatments.

According to archived web pages, the 7 April notice was reproduced on the website of the School of Information Science and Technology at Fudan University in Shanghai, but was subsequently removed. UK newspaper *The Observer* has reported that a similar notice was posted on, and then removed from, the website of the China University of Geosciences in Wuhan.

Helpful policies

Several researchers in China welcome the vetting process for COVID-19 studies. Alice Hughes, a conservation biologist at the Chinese Academy of Sciences (CAS) Xishuangbanna Tropical Botanical Garden, says the measure will stop the dissemination of potentially inaccurate and sensationalist research, such as a controversial study published in the *Journal of Medical Virology* in January, which suggested that snakes were the virus’s host.

Hughes says her institute’s director told her in late February that research on COVID-19 required MOST approval. She has not seen official policy documents herself. In early March, she says, she had a paper approved by the CAS and then by MOST within 72 hours. “We are continuing to see China publishing papers on the origins through this system,” she says.

Zhang Zhigang, an evolutionary microbiologist at Yunnan University in Kunming who published on the outbreak’s origins before the vetting process came in, also thinks it’s a good way to control research quality and reliability.

But news of the policies hasn’t reached all scientists or institutions. Chen Jin-Ping, an animal-disease researcher at the Guangdong Institute of Applied Biological Resources in Guangzhou who is also studying the virus’s origins, says he hasn’t been told that he needs ministry approval for his research to be published. And Fei Ma, dean of research and

graduate studies at Xi’an Jiaotong–Liverpool University in Suzhou, China, says he hasn’t heard of the need for coronavirus-related research to be approved by MOST or other government agencies.

Denis Simon, executive vice-chancellor at Duke Kunshan University, says his institute hasn’t received any official notices, but that researchers are discussing the issue.

Some researchers outside China fear the vetting process could hold up the release of important research. “Right now we desperately need all kinds of research relating to SARS-CoV-2, from basic studies to understand mechanisms of disease to vaccines and therapeutics,” says Ashley St. John, a virologist at the Duke–NUS Medical School in Singapore.

“We can’t afford any delays right now.”

Understanding the origin of SARS-CoV-2 could also lead to warning systems for virus spillovers from animals to people, she says.

Sarah Cobey, an infectious-disease researcher at the University of Chicago in Illinois, adds that it would be problematic if results from China were being filtered or suppressed for reasons other than quality. Observations of viral spread across countries inform the use of interventions such as social distancing, she says.

“If the research presents a biased picture, much of the record can eventually be corrected through studies of SARS-CoV-2 elsewhere,” she says, “but the distortion and delay would probably come at the cost of human health.”

COVID-19 COULD RUIN WEATHER FORECASTS AND CLIMATE RECORDS

As environmental-monitoring projects go dark, data that stretch back for decades are about to get gappy.

By Giuliana Viglione

Twice each year, Ed Dever’s group at Oregon State University in Corvallis heads out to sea off the Oregon and Washington coasts to refurbish and clean more than 100 delicate sensors that make up one segment of a US\$44-million-per-year scientific network called the Ocean Observatories Initiative. “If this had been a normal year, I would have been at sea right now,” he says.

Instead, Dever is one of many scientists sidelined by the coronavirus pandemic, watching from afar as precious field data disappear and instruments degrade. The scientific pause could imperil weather forecasts and threaten long-standing climate studies. In some cases, researchers are expecting gaps in data that have been collected regularly for decades. “The break in the scientific record is probably unprecedented,” says Frank Davis, an ecologist at the University of California, Santa Barbara.

Davis is executive director of the Long Term Ecological Research (LTER) programme, a

network of 30 sites stretching from the far north of Alaska all the way down to Antarctica. Consisting of both urban and rural locations, the LTER network allows scientists to study ecological processes over decades – from the impact of dwindling snowfalls on the mountains of Colorado to the effects of pollution in a Baltimore stream. At some sites, this might be the first interruption in more than 40 years, he says. “That’s painful for the scientists involved.”

Weather forecasting takes a hit

Other monitoring programmes are facing similar gaps. Scientists often ride along on the commercial container ships that criss-cross the world’s oceans, collecting data and deploying a variety of instruments that measure weather, as well as currents and other properties of the ocean. Most of those ships are still running, but travel restrictions mean that scientists are not allowed on board, says Justine Parks, a marine technician who manages one such programme at the Scripps Institution of Oceanography in La Jolla, California.

Port strikes and political instability have halted specific cruises in the past, Parks says. But, to her knowledge, this is the first time that the entire programme has shut down for an extended period of time.

Measurements made at sea are important for

“The break in the scientific record is probably unprecedented.”



Scientists are skipping trips meant to maintain sensors for the Ocean Observatories Initiative.

forecasting weather over the oceans, as well as for keeping longer-term records of ocean health and climate change, says Emma Heslop, a programme specialist in ocean observations at the Intergovernmental Oceanographic Commission in Paris. Her group is still trying to assess the extent of the damage that the pandemic is doing to the ocean-observing community as a whole, but researchers are already feeling some effects. Over the past two months, they've seen steadily declining numbers of shipboard observations – amounting, since the beginning of February, to a 15% loss of stations reporting data. And although the community is working hard to figure out other ways to collect important data, the situation is likely to worsen as the pandemic stretches on. “The longer the restrictions are in place,” Parks says, “the longer it will take for our operations to recover.”

Commercial flights provide invaluable weather data, too – measuring temperature, pressure and wind speeds. The meteorological data provided by the US aircraft fleet had decreased to half its normal levels as of 31 March, according to the US National Oceanic and Atmospheric Administration (NOAA).

Maintenance woes

Satellites and weather balloons can fill in some gaps, but certain aircraft data are irreplaceable. “It’s certainly the case that with the virtual loss of worldwide aviation, there is a gap in some of the records,” says Grahame Madge, a spokesperson for the UK Met Office in Exeter.

The Met Office estimates that the loss of aircraft observations will increase their forecast

error by 1–2%, but notes that, in areas where flights are typically more abundant, scientists’ forecast accuracy might suffer even more. The Met Office maintains more than 250 UK weather stations that provide continuous or daily feeds of autonomously collected atmospheric and weather data. For now, those systems are functioning just fine, but if an instrument goes down, Madge says, it will be difficult to get staff out to fix the problem.

Many of the world’s atmospheric-monitoring data are collected with little to no human intervention, and such projects should be able to keep running. The Advanced Global Atmospheric Gases Experiment, for example, measures ozone-depleting compounds, greenhouse gases and other trace components in the atmosphere at 13 remote sites around the globe. Many of their systems are autonomous: the stations are each staffed by one or two people who perform routine maintenance to keep the instruments running. Ray Weiss, an atmospheric chemist at Scripps who leads the project, says that two instruments have broken down so far, but the loss of a single instrument or even a whole site for a few weeks is unlikely to jeopardize the network’s monitoring capabilities. Arlyn Andrews, who runs NOAA’s greenhouse-gas-monitoring programme, says that impacts on that network have been “relatively minor”, and less than 5% of the NOAA sites have lost data so far.

Unless the situation gets a whole lot worse, Weiss anticipates that the programme will escape relatively unscathed. “We’re limping through, is the bottom line.”

Q&A



Charlie Swanton

Cancer-evolution researcher Charlie Swanton at the Francis Crick Institute in London has led the conversion of some labs into a coronavirus testing facilities. Swanton, also a consultant oncologist at the University College London Hospitals (UCLH), spoke to *Nature* about the effort.

How did this start?

Scientists didn’t want to sit at home and read reports about increases in deaths. We reached out to the UCLH and they said there was an unmet need for staff and patient testing. So researchers set up a working group to convert laboratories here into a rapid real-time polymerase chain reaction (RT-PCR) screening facility. Five large laboratories here have now been repurposed. Everybody wanted to help.

What does it take to retool a cancer lab into a diagnostic testing facility?

You need the right people, laboratory infrastructure and reagents. We have here BSL-3 (biosafety-level-3) facilities and BSL-3 trained staff, 10–15 RT-PCR machines, environments to extract RNA from viral samples, and space. We repurposed a lot of the software tools that we use to track patients’ cancer and blood samples to help us track COVID-19 tests.

We get swabs couriered from the UCLH every day: they’re taken up in an isolated coronavirus-specific lift and barcoded; the virus is inactivated, the PCR test done and the results reported back via a messaging app to medical staff. We are currently doing hundreds of tests per day, and hope to get up to 500–1,000 tests per day.

How have researchers adapted?

This is a new way of working for many of our scientists and staff. Much of this diagnostic work is repetitive and quite boring, but the stakes are high. It’s been extraordinary to see the selflessness of scientists here to help in the bigger effort of getting medics back to the front line.

Interview by Noah Baker

This interview has been edited for length and clarity.



India has been in lockdown because of the coronavirus pandemic since 25 March.

INDIA TRIES TO SLOW THE CORONAVIRUS WITHOUT WIDESPREAD TESTING

Country of 1.3 billion people deploys vast surveillance network to trace and quarantine those infected.

By Gayathri Vaidyanathan

Like many nations, India does not have enough kits to test most of its population for the new coronavirus. Instead, it is relying on people power: thousands of health-care workers are fanning out across the country to trace and quarantine people who might have had contact with those with COVID-19. People are typically tested only if they develop symptoms.

Countries such as South Korea isolated infected people on the basis of widespread testing, but some scientists say that India's mass-surveillance approach could achieve a similar goal, and be relevant for other low- and middle-income countries facing kit shortages.

Testing must still be part of India's COVID-19 strategy, and must be expanded rapidly, or infections will be missed, says Gagandeep Kang, executive director of the Translational Health Science and Technology Institute in Faridabad. Some people with the virus don't have symptoms, so their infections won't be detected otherwise. "There is going to be no solution to this without testing," says Kang. As of 19 April, the country had conducted some 400,000 tests – one of the lowest test rates per capita in the world. Still, epidemiologists say that India's strategy to trace and quarantine contacts, along with the government ordering the country's roughly 1.3 billion inhabitants

to stay at home for 21 days from late March – a deadline that has since been extended to 3 May – have probably helped to slow the spread of the virus in some places, and bought the country time to prepare its ailing health-care system.

Most people are allowed out for essentials, such as food and medical care, but in most states those under quarantine are closely monitored by social workers, and in some areas they can't leave their homes. The stakes are high: lockdowns are very hard on people who must work to feed themselves.

Contact tracing

India was well placed to send scores of public-health workers into villages, towns and cities to trace contacts and quarantine people. Its Integrated Disease Surveillance Programme (IDSP) already monitors people across the country for communicable diseases, and has been used to track H1N1 influenza and measles.

The network started watching for COVID-19 soon after the first case arrived in India in late January. When workers identify clusters of disease, a containment zone is set up and everyone inside it must stay at home. Social workers go door to door, and anyone with symptoms is tested, along with members of their household and close contacts. Those who test positive are taken to isolation units or hospitals.

Since monitoring began, the network has

helped to identify a majority of India's confirmed cases, which stood at 17,625 as of 20 April, and put hundreds of thousands of people under surveillance. (Just 543 people are known to have died of COVID-19 in India.)

Regional differences

The network is strongest in rural areas and urban slums, where people rely on government services, says Giridhara Babu, an epidemiologist at the Public Health Foundation of India in Bengaluru. The network's gaps are in upper-class urban areas where people have an incentive to hide their illness, fearing they will be ostracized if they test positive, he says.

The network is also stronger in some regions than others. Some states, such as Kerala, are obtaining phone records to investigate the contact histories of some people with COVID-19, says Amar Fettle, an epidemiologist with the state's IDSP in Thiruvananthapuram. In Chennai, a city of roughly seven million people in the state of Tamil Nadu, health workers are going door to door daily to monitor and test anyone showing signs of influenza-like illness.

Babu thinks it's too early to say whether mass contact tracing is working so far, especially in some states. Confirmed cases are low compared to many other countries, especially given India's huge population. But that could be a result of low testing rates, he says.

For the current strategy to work, it needs to capture all infections – a herculean task, says Ronojoy Adhikari, a mathematician at the University of Cambridge, UK. He co-authored a preprint that estimates the virus's spread in India with and without measures such as the lockdown (R. Singh and R. Adhikari. Preprint at <https://arxiv.org/abs/2003.12055>; 2020). His model, which has not been peer reviewed, says that if even 100 infected people escape detection and re-enter the population after the lockdown, which is highly likely, cases will resurge quickly. "That's really the crux of the matter here. How many people can we really effectively contact trace?" Adhikari says.

A long period of social distancing could help to keep infection levels manageable for the health-care system until a vaccine is developed, his model shows.

But there is a price. Strict social-distancing measures mean that people must stay at home, so many cannot work – particularly those on a daily wage. Developing nations do not have much financial flexibility to pay their inhabitants to stay at home for long, says Ricardo Hausmann, an economist at Harvard University in Cambridge, Massachusetts.

India has announced a 1.7-trillion-rupee (US\$22.6-billion) stimulus package for the poor, but economists have called it modest. India has to weigh the deaths that will be caused by the loss of livelihoods against those from the disease. "For those who have to stay at home, they starve to death," Hausmann says.

MANUNATH KIRAN/AFP/GETTY

HOW HOT WILL EARTH GET BY 2100?

Critics have challenged some assumptions behind global-warming studies. Researchers are now using a fresh set of scenarios to model the future of the planet.

By Jeff Tollefson

As world leaders gathered to mark the start of 2050, they looked back on the coronavirus pandemic 30 years before as a turning point in the quest to rein in global warming. Nations pulled together to defeat the pandemic, and that launched a new era of cooperation to prevent a climate disaster. Investments in green energy and new technology yielded rapid cuts in emissions of carbon dioxide, putting the world on track to limit global warming to around 1.5 °C above pre-industrial levels.

Or maybe not. In 2050, the world could look back and see the pandemic as little more than a blip in a long and mostly futile effort to stave off global warming. Despite a temporary drop in carbon emissions from the 2020 outbreak, countries turned to cheap fossil fuels to revive their economies after the crisis. Carbon emissions soared and temperatures followed, setting the stage for 5 °C of warming by the end of the century.

These are just two possible visions of the future. Nobody knows how the current pandemic will play out; nor is it clear whether humanity will ultimately come together to avoid a potential climate catastrophe. But climate researchers need to explore what kinds of problem might emerge with different levels of warming. So they have developed a suite of scenarios intended to represent a range of futures that humanity could face¹. Their goal is to investigate how different policies might alter carbon emissions – and how the planet will react to all of that heat-trapping gas.

At one end of the spectrum, optimistic scenarios explore worlds in which governments

join forces to advance low-carbon technologies while reducing poverty and inequality. The other end sees countries ramp up their use of cheap fossil fuels, pursuing economic growth at all costs.

Research teams have been running these scenarios through the world's major climate models for the first time, providing projections of how Earth might respond to different socio-economic pathways. These simulations will inform climate research for years to come, and will play a central part in the next major assessment of global warming by the Intergovernmental Panel on Climate Change (IPCC), which is due out next year. The research could also have a key role in the negotiations around a new set of commitments to reduce emissions under the 2015 Paris climate agreement.

These scenarios update a set that has been in use for the past decade, including one extreme – and controversial – version that projects a temperature increase of around 5 °C above pre-industrial levels by 2100. Critics have charged that this particular scenario, which has had a central role in climate studies for more than a decade, is misleading because it includes unrealistic amounts of coal use – a roughly fivefold increase by 2100. But many researchers dismiss that criticism, saying that even such high-emissions scenarios have value as long as people understand their underlying assumptions and limitations. A massive release of methane from Arctic permafrost, for example, could have a similar effect to huge surges in fossil-fuel use.

“We’re trying to understand risks, not predict the future,” says Donald Wuebbles, an atmospheric scientist at the University of Illinois at Urbana–Champaign and a





coordinating lead author on the first volume of the latest US national climate assessment², released in 2017. The scenarios are not designed to project emissions, but to investigate different levels of warming and types of economic development. They help a wide variety of researchers: climate modellers use them to test their models and project the impact of increasing greenhouse-gas emissions; economists need them to explore the costs of policies; and ecologists rely on them to predict changes to ecosystems around the globe.

“This is not science fiction,” says Kristie Ebi, an environmental-health researcher at the University of Washington in Seattle who co-chairs the committee that developed the new scenarios. “We need these model results to give us insights into the impacts of our choices, and now we can do that.”

Unusual business

In April 1989, a group of experts tasked with forecasting potential futures met in Bilthoven, the Netherlands, to prepare for the first IPCC assessment, which was due out the following year. They created scenarios describing how much carbon dioxide, methane and other heat-trapping gases nations might produce over the next century³. And those possible future worlds – from the extremely polluted to the exceptionally clean – provided the raw material for climate modellers to project how the planet might react.

Since then, the IPCC has updated the main emissions scenarios several times. But the situation changed in 2006, when the IPCC decided to get out of the scenario-development

“We’re trying to understand risks, not predict the future.”

business because of pressure from the United States and others who argued that the organization should assess, not guide, science.

So, in 2010, a self-appointed group led by climate scientist Richard Moss, then at the Joint Global Change Research Institute in College Park, Maryland, published a new framework for creating and using scenarios designed to guide research for the IPCC’s last assessment⁴, which was released in 2013–14.

The group provided a set of four projections of future carbon pollution levels – dubbed Representative Concentration Pathways (RCPs) – that could be run by climate-modelling groups around the world to produce forecasts about the fate of the planet⁵. The RCPs were selected to portray different levels of radiative forcing – a number that reflects how much extra warming results from greenhouse-gas emissions. The

RCPs weren't intended to describe particular emissions trends or project how economies and technology might change. That job was left for other researchers, who would later produce sets of emissions trends that could drive greenhouse-gas concentrations in ways that mimic the RCPs.

Moss says the RCPs were designed to capture the spectrum of warming possibilities in the scientific literature and create a significant enough range between the high and low projections that climate modellers would be able to differentiate between them. And one major appeal of the scenario with a 5 °C global temperature increase that elicited so much criticism – called RCP8.5 – is that it provides modellers with a powerful signal. “We wanted to give enough detail so that climate modellers could do their work,” says Moss. Regarding the individual scenarios, he adds, “we never meant to give them any particular weight”.

Over time, however, the RCPs took on a life of their own. Although the caveats and qualifications are all there for those who know where to look, many scientists and others started using RCP8.5 to represent a world without aggressive climate action.

“It's very tempting to use RCP8.5 for a whole range of reasons, but it's also pretty unrealistic,” says Glen Peters, a climate-policy researcher at the Center for International Climate Research in Oslo and co-author of a recent commentary on the issue⁶. “The question is how you balance those issues and communicate what it represents.”

The mischaracterization of RCP8.5 – as a projection of what could happen in a business-as-usual world in which governments fail to enact climate policies – is endemic, says Roger Pielke Jr, a science-policy researcher at the University of Colorado Boulder. Pielke says that even major scientific reviews such as the US national climate assessment have defaulted to using RCP8.5 as a de facto baseline scenario in which emissions continue to spike. That inflates projections of the effects of global warming – as well as of the costs of inaction, he says.

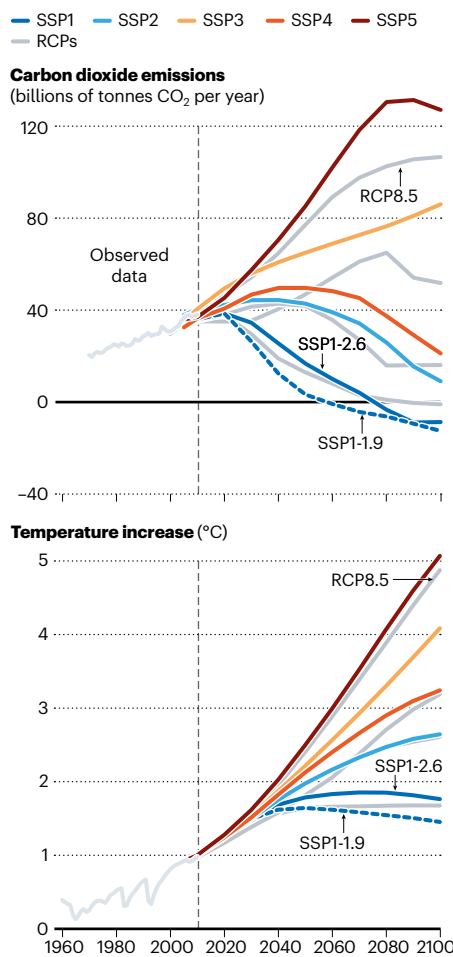
Wuebbles defends the decision to use RCP8.5 in the US assessment². The document refers to RCP8.5 merely as a “higher” scenario. It notes that emissions were consistent with this scenario for 15–20 years, until they levelled off for a few years around 2014.

Moreover, RCP8.5 provides scientists with a high-risk scenario that is valuable for understanding the risks posed by climate extremes, says Céline Guivarch, a climate-change economist at the Centre for International Research on the Environment and Development (CIRED) in Nogent-sur-Marne, France. Many scientists argue that even if coal use doesn't rise in a catastrophic way, 5 °C of warming could occur by other means, including thawing permafrost.

After the RCPs were published in 2010,

A RANGE OF FUTURES

Researchers have developed new scenarios, called Shared Socioeconomic Pathways (SSPs), to explore different ranges of development and how they would alter the climate. These complement older scenarios called Representative Concentration Pathways (RCPs).



the plan was to have a new set of fleshed-out socio-economic scenarios ready within two years. Those would have fed into the IPCC reports that came out in 2013 and 2014, which found that the rate of warming since 1950 is unprecedented over a timescale of centuries to millennia, and set the stage for the 2015 Paris climate accord.

But the process was much more difficult – and took a lot longer – than anticipated. The new generation of scenarios, known as Shared Socioeconomic Pathways (SSPs), were not introduced until 2015. Only now, as the major climate-modelling centres around the world run their experiments for the 2021 IPCC assessment, are they taking centre stage in climate research.

Although based on the old RCPs, the new scenarios for the first time present fully fleshed-out narratives about how the world might evolve (see ‘A range of futures’). Each provides a broad storyline about how the world might change, as well as numbers for key demographic trends – population, economic productivity, urbanization and education – in every country on Earth, which modellers then use to simulate

emissions and planetary impacts.

The teams that produced the SSPs intentionally left out any climate policies. This approach allows scientists to run their own experiments and test the impacts of different decisions by governments and societies, says Ebi. The flexibility allows her and other public-health researchers to compare and contrast the health benefits from climate policies that simultaneously reduce carbon emissions and result in cleaner air.

“You couldn't do that before,” Ebi says. “It's allowing the climate community to ask questions that we couldn't ask.”

Rocky road

Although the SSP scenarios are only a few years old, they were developed in a world very different from today's. They were shaped before the political upheaval of 2016, when the United Kingdom voted to exit the European Union and the United States elected President Donald Trump, who promised to put America first and withdraw from the Paris climate treaty.

But the teams that drafted the SSPs imagined a storyline that is very close to the path that the United States and other major powers are taking. The SSP3 scenario, called “regional rivalry – a rocky road”, is defined by a resurgence of nationalism. It sees concerns about economic competitiveness and security lead to trade wars. As the decades progress, national efforts to lock down energy and food supplies short-circuit global development. Investments in education and technology decline. Curbing greenhouse gases would be difficult in such a world, and adapting to climate change wouldn't be any easier. Under this scenario, the average global temperature is projected to soar to more than 4 °C above pre-industrial levels.

For Ebi, it's a lesson in humility, because the scenario seemed outlandish when it was developed. But that is the point.

“When we started working on this, there was no discussion of America first, there was no Brexit, there weren't trade wars between the United States and China,” she says. “It's uncomfortable, but you need to have those kinds of pathway. We don't know what the future is going to look like.”

Jeff Tollefson is a reporter for *Nature* in New York.

1. O'Neill, B. C. et al. *Glob. Environ. Change* **42**, 169–180 (2017).
2. US Global Change Research Program. *Fourth National Climate Assessment, Vols I–II* (US Global Change Research Program, 2017–18).
3. Intergovernmental Panel on Climate Change. *Climate Change: The IPCC Scientific Assessment* [IPCC First Assessment Report] (Cambridge Univ. Press, 1990).
4. Intergovernmental Panel on Climate Change. *Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (IPCC, 2014).
5. Moss, R. H. et al. *Nature* **463**, 747–756 (2010).
6. Hausfather, Z. & Peters, G. P. *Nature* **577**, 618–620 (2020).

SOURCE: ADAPTED FROM FIG. 3. B. C. O'NEILL ET AL. *GEOSCI. MODEL DEV.* **9**, 3461–3482 (2016)

MEASLES IS ON THE RISE — AND COVID-19 COULD MAKE IT WORSE

Measles has killed thousands in the Democratic Republic of the Congo and is surging worldwide. But more than 20 countries have suspended vaccination campaigns to cope with the new coronavirus threat. **By Leslie Roberts**



A driver prepares to transport measles vaccines in the Democratic Republic of the Congo.

A viral outbreak has killed more than 6,500 children in the Democratic Republic of the Congo (DRC) and is still spreading through the country. The foe isn't the feared coronavirus, which has only recently reached the DRC. It's an old, familiar and underestimated adversary: measles.

Cases began to spike here in October 2018. Children became weak, feverish and congested, with red eyes and painful sores in their mouths, all with the telltale rash of measles. "We have been running after the virus ever since," says Balcha Masresha, an epidemiologist with the World Health Organization (WHO) regional Africa office in Brazzaville in the neighbouring Republic of Congo. The situation has mushroomed into what WHO experts say might be the largest documented measles outbreak in one country since the world gained a measles vaccine in 1963 (see 'Measles cases on the rise').

Despite the vaccine, the highly contagious measles virus continues to spread around the globe. In 2018, cases surged to an estimated 10 million worldwide, with 140,000 deaths, a 58% increase since 2016. In rich countries,

scattered measles outbreaks are fuelled by people refusing to vaccinate their children. But in poor countries, the problems are health systems so broken and underfunded that it is almost impossible to deliver the vaccine to people who need it. The DRC's flood of cases shows why measles will keep flaring up despite efforts to control it. And the situation will worsen with the COVID-19 pandemic: 24 countries had suspended measles vaccination campaigns by 14 April as health-care workers scrambled to deal with the coronavirus.

An overlooked killer

In poor countries, measles is a killer, especially when combined with malnutrition and vitamin A deficiency. Estimates are uncertain, but the death rate in developing countries hovers around 3–6%, and it can spike as high as 30%, the WHO says. Its victims often die of complications including pneumonia or diarrhoea and dehydration. Others can be left with permanent disabilities, including blindness, hearing loss and brain damage. The virus also impairs the immune system for months or years after infection, creating 'immune amnesia' that leaves children vulnerable to other infections.

The virus is so contagious that few unvaccinated people who come into contact with it are spared its effects. Scientists define infectiousness using the 'reproduction number' — how many people, on average, would be infected by a single person with the virus, in a population that has no immunity. For Ebola, that number is estimated at 1.5–2.5. The new coronavirus terrifying the world seems to be somewhere between 2 and 3. Measles tops the charts with a reproduction number of 12–18, which makes it the most contagious virus known. You don't need to be in the same room as an infected person to catch the virus — it is spread by respiratory droplets that can linger in the air for hours.

Two doses of a safe and effective vaccine can prevent measles. Many children in poor countries are lucky to get a single dose, however, which doesn't always lead to full protection in all who receive it. Because the virus is so contagious, 92–95% of a population needs to be fully immunized to ward off outbreaks. In the DRC, only 57% of children received even one dose of measles vaccine in 2018, according to a Unicef study (go.nature.com/3cfjalh), creating ideal conditions for the virus to explode.

Failure modes

In other countries, too, measles simmers at low levels until the number of children susceptible to the virus builds up and it takes off. In each country, a slightly different mix of factors leads to an outbreak.

In Ukraine, after a child died of unrelated causes following a measles jab in 2008, vaccination coverage plummeted from 95% that year to 31% in 2016, says Robb Linkins, a measles specialist in the global immunization division at the US Centers for Disease Control and Prevention (CDC) in Atlanta, Georgia. He says no one was surprised when, in 2017, a huge outbreak hit that has led to more than 115,000 cases.

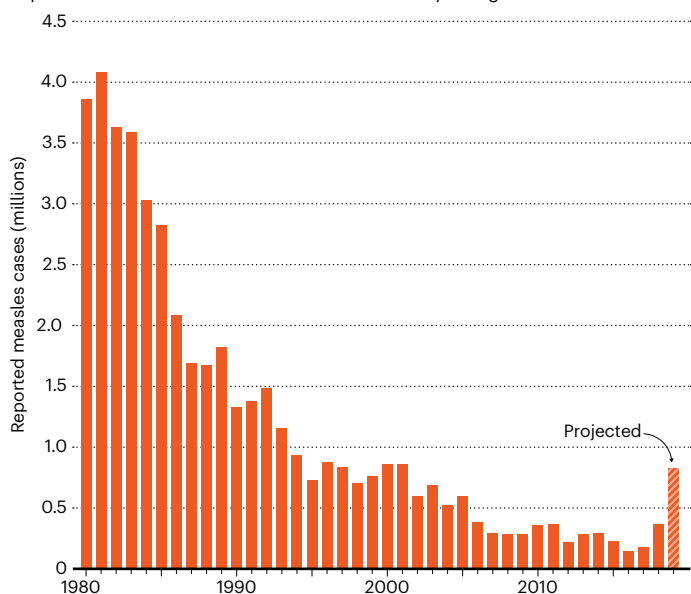
In Madagascar, a shortage of measles vaccine helped fuel an outbreak that has swept the island nation starting in 2018, causing more than 240,000 cases and 1,000 deaths.

The DRC has a number of difficulties. The country has such a high birth rate — 3.5 million children born each year — that it needs to conduct mass vaccination campaigns every two years. Those campaigns, in which tens of thousands of health workers fan out across this vast country, are a logistical nightmare. The vaccine must travel from the capital, Kinshasa, to remote villages that can be reached only by helicopter — or through bloody conflict zones.

The vaccine must be kept at between 2 °C and 8 °C from the time it leaves the warehouse until it is administered — a challenge in a tropical environment where power cuts are frequent. Health workers must be trained to inject it safely. The vaccine comes as a powder, which must be reconstituted with sterile diluent and

MEASLES CASES ON THE RISE

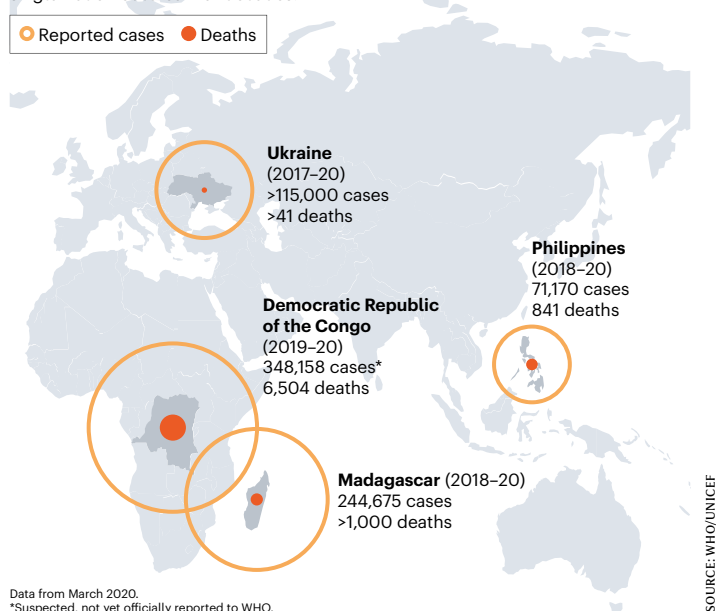
Reported measles cases are estimated to be at a 20-year high.



*Reported measles cases massively undercount true measles cases (estimated at ten million for 2018), but signify trends.

LARGE MEASLES OUTBREAKS

The epidemic in the Democratic Republic of the Congo is the largest single-nation outbreak for decades.



SOURCE: WHO/UNICEF

used within 6 hours. It also comes in ten-dose vials; worried about wastage, vaccinators are sometimes hesitant to open one when just a few children show up to a session, so children go unimmunized. Clinics also have to be open when parents can make it, and vaccinators have to be paid or they won't come, either – a problem in a country plagued with corruption.

The DRC has also been battling Ebola, outbreaks of cholera and yellow fever – and now the coronavirus. Measles often takes lower priority. In addition, the DRC “is confronted not only with political challenges but a long-running civil war”, says Katrina Kretsinger, a medical epidemiologist and global measles expert at the WHO in Geneva, Switzerland.

Money is a major problem. Vaccination campaigns cost around US\$1.80 per child in the DRC, says Masresha; international donors foot only part of the bill. In 2010, the DRC couldn't muster enough funds and cancelled a scheduled campaign. An outbreak that hit at the end of the year raged for more than 30 months. Further campaigns in 2013–14 and 2016–17 didn't reach enough children. In June 2019, after cases soared to more than 3,500 a week at the start of the year, the DRC government declared an epidemic, opening the door to further international aid. By the end of the year, 18.5 million children had been vaccinated.

The WHO estimates that there have been more than 348,000 cases and 6,500 deaths, but Francisco Luquero, an epidemiologist at Epicentre, the research arm of Médecins Sans Frontières (MSF, also called Doctors Without Borders) in Paris, thinks the outbreak is much worse. The case count reflects only people who go to health centres, he says; many don't in the DRC. As for the mortality estimates, “they count deaths that happen right after

a measles case. They should look out for the next five years,” he says, because of immune amnesia. “The outbreak will have a profound impact on public health.”

Eradication hopes

Steve Cochi, a paediatrician and senior adviser to the CDC's global-immunization division, is especially frustrated by measles's global toll because, from a biological and technical standpoint, he says, the disease could be eradicated. Unlike Ebola, yellow fever or (probably) the new coronavirus, it has no animal host, and a cheap and effective vaccine exists.

A new vaccine delivery system being developed by two teams – one a collaboration between the CDC and the Georgia Institute of Technology and Micron Biomedical, both in Atlanta, and the other at Vaxxas, a biotech company based in Sydney, Australia – could be a “game changer” for measles control, says Cochi. The vaccine uses a microarray patch, which looks like a small, round bandage with hundreds of microneedles, each carrying a small amount of live, freeze-dried vaccine. It delivers vaccine under the skin in 5 minutes.

“It is very thermostable, takes up little space, doesn't have to be reconstituted, and you don't have to worry about safety,” Cochi says. But the measles patch has languished for lack of funding, he says, and has yet to reach clinical trials.

Until polio is eradicated, the world does not have the appetite or money to target another disease for extinction, Cochi says. In 2010, the WHO's key Strategic Advisory Group of Experts on Immunization (SAGE) declared that measles can and should be eradicated, but didn't recommend a target date. Since then, advocates have been lobbying the WHO to launch a global measles-eradication campaign

and set a date for completion, as it did for smallpox and polio. At a meeting last October, however, SAGE recommended a different tack: waiting until success is in sight – say, five years away – before pushing full-bore to wipe out the disease. Doing so would require boosting rates of routine immunization with two doses of measles vaccine to a level never achieved before. The DRC is one of about 20 countries that have yet to add the second dose to their regimes. Eradication would also depend on improving the quality of mass campaigns and bringing an improved vaccine into use.

Coronavirus crunch

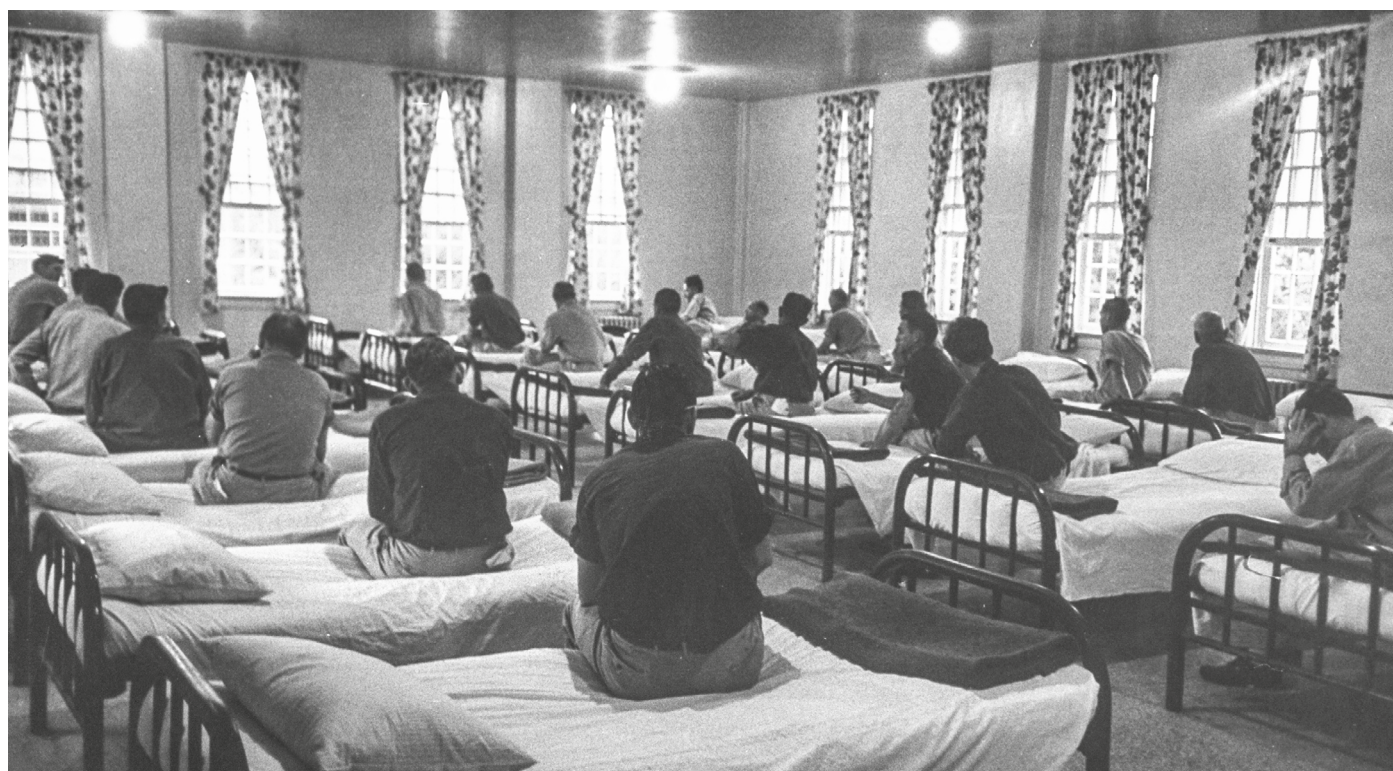
The coronavirus pandemic has dealt measles-control efforts another huge blow. On 26 March, SAGE recommended that countries suspend all preventive mass-vaccination campaigns, including for measles. Already, 37 countries have suspended scheduled measles campaigns or will soon do so, says Linkins. This means that 117 million children may not be vaccinated as planned, he says. The DRC, however, is continuing its outbreak response.

“We must protect vulnerable populations from the spread of COVID-19,” Linkins says, but limiting preventive measles immunization will create “dangerous immunity gaps”. Countries must be able to resume their campaigns quickly after the pandemic subsides, he adds.

With campaigns cancelled and global measles-immunization rates for just one dose of vaccine stalled at 86%, the unrelenting cycle of outbreaks will continue. This DRC outbreak will subside, but Masresha says it will be a “temporary victory”: the virus will rebound.

Leslie Roberts is a science reporter in Washington DC.

Books & arts



Patients at Milledgeville State Hospital, Georgia, in 1951.

Psychiatry under the shadow of white supremacy

From the start, racism has shaped the care of people with mental illness in the United States. **By Mical Raz**

How does a culture that enslaved people, encouraged lynching and developed racial segregation decide who is and is not sane? That is the question that frames Mab Segrest's book on the legacy of slavery for US psychiatry in general and for what was in the 1940s and 1950s one of the largest psychiatric hospitals in the world.

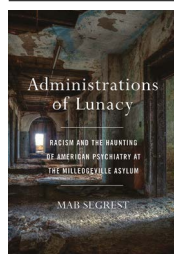
Combining archival research with fictionalized scenes, Segrest, a feminist and anti-racist scholar, recounts more than a century of custodial care at Georgia's infamous Milledgeville hospital for people with mental illness. It opened in 1842 and Segrest weaves its history with the wider trajectory of US psychiatric care, the ravages of the American Civil War (1861–65) and the many manifestations of

white supremacy and violence against women.

The book is organized chronologically, but includes multiple forays into the present, which can be distracting. Segrest begins her account by reconstructing a period when slavery was omnipresent and the asylum took in only white patients. After the war,

psychiatry followed a racist trajectory that was by no means inevitable, argues Segrest. When Milledgeville started taking black patients in 1867, it – like other asylums nationwide – adopted racial segregation.

As part of their treatment, white men worked as gardeners; black men had to labour on the institution's farm. White women were seamstresses; women of colour worked in the laundry. Segrest uses the asylum's archive to show that luxuries such as writing supplies, slippers, soap and carpets were allocated much more generously to white patients, whereas black patients faced daily discrimination and neglect. Many died soon after arrival, reflecting both their poor health and the deplorable conditions they had to endure.



Administrations of Lunacy: Racism and the Haunting of American Psychiatry at the Milledgeville Asylum
Mab Segrest
The New Press (2020)

Segrest also highlights what was left unexplored. Rather than asking how slavery might devastate an individual's psyche, physicians treating newly freed African Americans discussed how their mental health might have been harmed by emancipation. What's more, these patients often came from counties in which extreme racial violence, including "whippings, assaults, and murders", was routine. Yet, in many cases, this history remained undocumented. Asylum psychiatry "maintained a vast silence about the bloodbath all around it", writes Segrest, just as it had previously been silent about the violence of slavery.

Lingering effects

Milledgeville underwent several name changes and ultimately became the Central State Hospital before the main building closed in 2010. In her final chapters, Segrest examines how, when such hospitals began to close in the 1980s, penal institutions took their place. As welfare programmes were starved, the US prison population spiked, with people of colour and people with mental illness disproportionately incarcerated. Today, 90% of US psychiatric-care beds are in jails and prisons. Psychiatry will not be able to escape "the after-life of slavery", she argues, until it confronts its culpability in mass incarceration.

A newcomer to the history of psychiatry, Segrest's approach is fresh and creative. She uses her imagination to flesh out the realities of life within the asylum walls. Describing Frances Edwards, a mother of seven taken to Milledgeville in 1856, Segrest imagines her arms feeling weirdly light and empty without her children, as her breasts "ached and leaked". Segrest also finds connections between topics not always identified as part of psychiatry's past. She calls attention to the high rates of infant mortality in the black community, exploring how such factors might have shaped – and still shape – black women's mental health.

Segrest's is one of several books in the past few years that have foregrounded discussions of race in the history of psychiatry and of asylums. Her impressionistic style and convoluted structure contrast sharply with the more rigorous work of historians such as Martin Summers in his 2019 *Madness in the City of Magnificent Intentions* and Wendy Gonaver in *The Peculiar Institution and the Making of Modern Psychiatry, 1840–1880* (2019). Segrest's mixture of fact and fiction can also be confusing.

But what is lost in clarity is perhaps gained in popular appeal. Uncomfortable reading at times, this valuable book helps to show how white supremacy shaped the definition and care of people with mental illness from the start, and how psychiatry remains in its shadow.

Mical Raz is a physician and historian of health policy at the University of Rochester, New York. e-mail: micalraz@rochester.edu

Preppers, bunkers and emaciated polar bears

How to live in the face of death – Mark O'Connell's personal journey. **By Caspar Henderson**

Are we facing the end of civilization, or even the planet? It's a question that attracted some serious scientific firepower even before the current pandemic. UK institutions such as the Centre for Existential Risk at the University of Cambridge and the Future of Humanity Institute at the University of Oxford are modelling the probabilities of various catastrophes, from a giant meteorite strike to a scenario in which criminals and psychopaths gain 'easy nukes' and incinerate a vulnerable world. Meanwhile, climate and Earth-systems scientists are amassing more evidence by the month that, barring rapid and profound reorganization in our societies, climate change will batter our world on at least the scale of a major war.

Rather than assessing the science itself, *Notes From An Apocalypse* explores how such threats affect individuals. Written before the COVID-19 crisis, it is an eerily prescient mix of confession, political critique, meditation and comic monologue on living in the face of death. It is the second such book from Mark O'Connell, the winner of the 2018 Wellcome Book Prize (for his first, *To Be A Machine*, which

tackled the philosophy that humanity can evolve beyond its limitations using science). As the scientific and political responses (or lack thereof) to threats ranging from global heating to plastic dominate the headlines, O'Connell probes deeper into our personal psyches. In a tone somewhere between those of writer Samuel Beckett, film-maker Woody Allen and poet W. B. Yeats, he asks what happens when we're faced with the prospect of both individual and global demise.

A successful literary journalist living in Dublin with his young family, O'Connell is obsessed with doom. He sets his computer home page to an online forum dedicated to discussing civilizational collapse, and compulsively checks his smartphone for YouTube clips of emaciated polar bears, when he should instead be watching cartoons with his son.

This fixation leads him into the shadowy worlds of 'preppers'. These self-styled survivalists stockpile stores and weapons, readying themselves for civilization's impending collapse, and feed endless online discussions about videos of the contents of their 'bug-out bags' – knapsacks containing items they



This converted nuclear missile vault in Glasco, Kansas, has a heated pool and water slide.

CHET STRANGE/THE NEW YORK TIMES/EYEVINE



Many former military bunkers, such as this one near Edgemont, South Dakota, are being repurposed into doomsday communities.

consider essential for the end-time. The clips strike O'Connell as apocalyptic variations of 'haul videos', in which young consumers lay out the treasures of a recent shopping binge, with Kevlar socks in place of Superga shoes, and athleisure swapped for military-grade cordage.

O'Connell hits the road, deploying his considerable gonzo journalism skills to seek out other doomsday obsessives, each caught up in their own dark, imagined futures. He visits high-end condos being built in a former weapons-storage facility in South Dakota that can withstand explosions of up to half a megatonne. He seeks out the luxury bolthole of Peter Thiel, billionaire entrepreneur and 'sovereign individual' – a person who controls vast resources and intends to redesign the government to suit their needs after collapse. Thiel is one of several Silicon Valley elites who have chosen to build their bunkers in New Zealand. And O'Connell attends a meeting of the Mars Society in Pasadena, California, where enthusiasts share dreams of a new, American-style frontier in the unspeakably harsh conditions on the red planet.

In each case, O'Connell skewers what he sees as a central psychopathology or distorted value system, even as he acknowledges his own uneasy fascination and near-complicity.

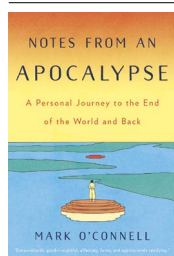
Preppers, he argues, are readying not for their fears but their fantasies; they contribute "nothing to the prevention or alleviation of suffering in others". Thiel and the would-be Mars colonizers imagine a world somehow beyond politics – and taxes – that is almost exclusively white and male.

The book's final two journeys explore a quieter response to the threat of catastrophe. O'Connell retreats to the Scottish Highlands with members of the Dark Mountain movement – artistic and mostly gentle souls who, knowing climate chaos is imminent, seek solace in reconnection with nature. Lastly, he takes part in an anti-stag party with a friend who is separating from his wife. They visit the Chernobyl Exclusion Zone in Ukraine, where a future imagined by the Soviets met its end. O'Connell is struck by the shards of "our own machine age" – shattered glass from broken

screens and a heap of old television sets with "ancient circuit boards greened with algae". He wonders if this might be a glimpse of our own future.

Back in Dublin, O'Connell finds he has lost his taste for cosmic nihilism. In the radiance, joy and hilarity of his kids – the way in which they connect with, rather than retreat from, the world – he finds inspiration to shift his focus from how our lives might end, to what makes them worth living.

Notes from an Apocalypse offers no scientific analysis of the existential threats we face or how we should respond. Instead, it illuminates the anxieties and delusions we share and oversights we commit, and shows how easily our fears (particularly when enabled by power, money and technology) can cause us to walk away from the disasters we create – to hide, flee, stockpile – just when we most need to engage. In this reflective, hilarious and disturbing page-turner, O'Connell makes a compelling case that connecting with nature and each other is the best way to calm our apocalyptic dread – and it might even increase our prospects of avoiding the worst.



Notes from an Apocalypse: A Personal Journey to the End of the World and Back

Mark O'Connell
Doubleday (2020)

Caspar Henderson is a writer and journalist in Oxford, UK.
e-mail: caspar81@gmail.com

Comment



A petrol station in Berlin. On average, drivers accurately judge fuel costs but severely underestimate all other expenditure.

Running a car costs much more than people think – stalling the uptake of green travel

Mark A. Andor, Andreas Gerster, Kenneth T. Gillingham and Marco Horvath

Car owners underestimate total vehicle costs. Giving consumers this information could encourage the switch to cleaner transport and reduce emissions.

Priate cars are responsible for about 11% of the world's total carbon dioxide emissions. That's the greatest share in the transport sector, which accounts for 24% of emissions overall¹. Petrol and diesel cars are associated with many other harmful effects, such as air pollution, congestion and accidents. It is clear that these cars must be largely removed from the roads to achieve sustainable mobility.

The good news is that some policies have

been enacted to reduce greenhouse-gas emissions and air pollution from petrol and diesel cars. Some markets have tightened emissions limits. In the European Union from 2021 onwards, for instance, the fleet-wide average emission target for new cars will be reduced by more than 25% – from 130 grams of CO₂ per kilometre (g CO₂ km⁻¹) in 2015–19 to 95 g CO₂ km⁻¹. The current average is 120 g CO₂ km⁻¹ (on the basis of 2018 data; see go.nature.com/39puqyy). And the United Kingdom will stop sales of new petrol, diesel and hybrid cars from 2035 onwards.

Cities are taking action, too. For example, Oslo has decreased the number of parking spots and raised parking fees². New York City and Shenzhen in China are electrifying their bus fleets, and London is reducing bus emissions and improving infrastructure for walking and biking. Stuttgart in Germany is among the cities banning older, polluting diesel cars.

The bad news is that more than 99% of new

passenger cars sold worldwide still rely on fossil fuels^{1,3}, and overall vehicle ownership in Europe grew by 25% between 2000 and 2017 (ref. 4). The continued demand for vehicle ownership stems from several factors, including increased income⁵ and more mobility as people travel farther to their jobs as a result of greater city sprawl⁶. The transition away from conventional vehicles is hindered by the high upfront costs of electric cars^{7–9} and too few charging stations, leading to 'range anxiety' from potential owners^{10,11}.

Consumers decide whether to own a vehicle on the basis of considerations such as where they live and the vehicle's upfront and lifetime costs¹². If they systematically underestimate total costs, this could increase car ownership and its associated emissions. It could also make alternative forms of transport – car sharing, alternative-fuel vehicles, public transport, biking or walking, say – seem less attractive.

We surveyed more than 6,000 citizens across

Comment

Germany to investigate whether consumers grasp the total cost of car ownership. We also performed a simple analysis to explore the potential implications of this awareness on the number of cars on the road.

We find that people underestimate the total cost of owning a car by about 50%. We also found that providing personalized information on the costs of car ownership increased respondents' willingness to pay for a public-transport ticket by around 22% (see Supplementary information; SI). We estimate that educating people in Germany about the true cost could reduce car ownership by up to 37% and cut associated transport emissions by 23%. Here, we suggest labelling and communication policies that could help to speed the transition to cleaner transport.

Data set and methods

We conducted a survey of the heads of German households – the people who self-report as being responsible for financial decisions – between 23 April and 12 June 2018. For every car owner, the survey elicited responses on the cost of ownership, based on the individual's car type and driving behaviour, as well as socio-economic characteristics such as income, number of children and education. The work was done in collaboration with the survey institute Forsa in Berlin. It used a random sample of Forsa's household panel, which is representative of the German-speaking population aged 14 and older.

Of the 7,823 individuals who started the survey, 6,812 completed it, 6,233 of whom own a car (92%). Of these, 5,483 stated what they thought their monthly car costs were. These respondents form the basis for our analyses. Taking households' car type and travel behaviour into account, we use detailed information from the German Automobile Club (ADAC) and other sources to calculate the actual monthly costs of car ownership, on average, for depreciation, fuel, taxes and insurance, and repair (see SI).

Striking difference

Our findings were striking. Consumers underestimate the total cost of vehicle ownership by €221 (US\$240) per month on average. The misjudgement amounts to 52% of the actual costs, so the total cost is nearly twice what people think. Using only the respondents who provided an estimate for all cost factors, the underestimation is €161 on average, which is 35% of the actual costs. To be conservative, we proceed in our analysis using this sample (for the full sample, see SI).

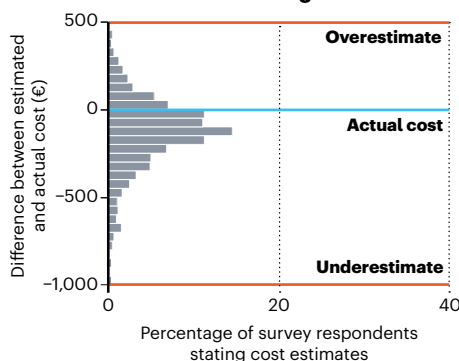
The difference between estimates and actual costs varies widely. This is not necessarily surprising. The mean and median of the difference is well below zero, clearly demonstrating that costs are underestimated on average.

We also investigated the four main costs of car ownership: fuel, depreciation, repair, and tax and insurance. On average, respondents came

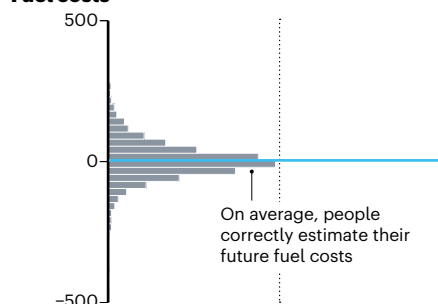
COSTLY MISJUDGEMENT

Car owners in Germany severely underestimated major forms of expenditure, except on fuel. Knowledge about true total costs could result in 37% fewer cars on German roads, an analysis finds. New transport policies should make such information available at the point of sale.

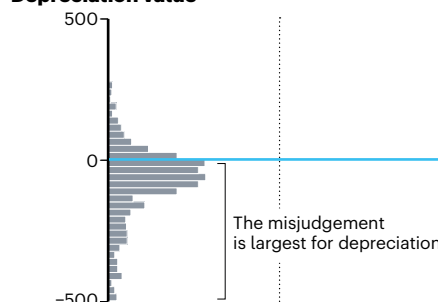
Overall actual costs of running a car



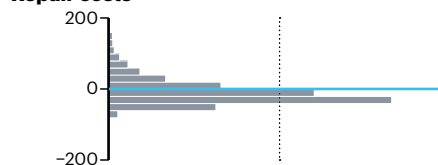
Fuel costs



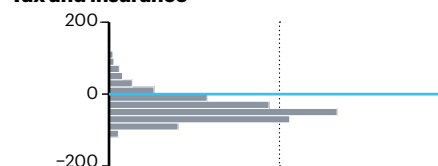
Depreciation value



Repair costs



Tax and insurance



very close to perfectly estimating how much they spent on fuel, consistent with the previous literature¹³. But they severely underestimated all other major expenditure for running their cars (see 'Costly misjudgement'). To our knowledge, this misjudgement has not been reported

previously, and it can provide leverage points for designing new transport policies.

Fewer cars

Let's assume that it is possible to completely eliminate the degree to which people systematically underestimate the total cost of owning a car. Would this change the number of cars on the road? We explored this question by modelling how car ownership changes when the associated costs change, on the basis of previous work on US car ownership¹². This allowed us to calculate the reduction in the 47.1 million passenger cars in Germany that one might expect if households were perfectly informed about running costs.

We predict that being aware of the true cost of owning a car could result in almost 17.6 million (37%) fewer vehicles on the road in Germany. Such a drastic reduction would mean less congestion and cleaner air. It would also lead to a drop in CO₂ emissions of about 37 million tonnes per year: 4.3% of Germany's total, or 23% of emissions from its transportation sector (see SI).

Although increased demand for public transport could lead to more CO₂ emissions from bus and rail travel, this effect would probably be small. First, the carbon emissions per person for each kilometre travelled for these modes of transport are around half those of car travel (see go.nature.com/2upeejh). Second, emissions-trading schemes in the EU, for example, mean that the increased electricity consumed by the growing number of electric trains and vehicles is prevented from translating into extra carbon emissions for the economy, because total emissions are capped.

Even a more cautious estimate of changes in car ownership in response to higher prices of new vehicles (as reported in a 2018 US impact analysis¹⁴) would still imply a 9% reduction in car ownership, taking more than 4 million cars off German roads (see SI).

We further used our survey data and empirical estimates of the impact of car-ownership costs to investigate whether such misjudgements affect the use of public transport and electric vehicles. We predict an increase in demand for bus and rail travel of 8% and 12%, respectively. Purchases of electric vehicles could increase by about 73% (see SI).

Although our survey was conducted in Germany, we expect the results to be applicable throughout Europe, and probably to countries with similar economies elsewhere. In 2017, Germany had 561 passenger cars for every 1,000 inhabitants; the EU average in the same year was 512 (ref. 15). And people in other countries accurately estimate fuel costs¹³, as in our survey.

These are preliminary findings. Our calculations require a number of assumptions. For example, we had to assume how much car ownership changes when household perceptions of total costs change, and how much

the demand for other transport modes would respond to decreased car ownership (see SI, pages 5–6). To explain, consider the case in which the changes in car ownership and the demand for other transport modes were less pronounced. Then, eliminating cost beliefs would have smaller effects.

Critics might argue that cost is merely one of many factors that influence individuals' decisions to own a combustion-engine car, including status, the need for mobility in rural areas and the lack of infrastructure for electric-vehicle charging and for public transport. Although cost is indeed only one factor, it is a crucial one in the car-ownership decision^{12,16}. Another potential criticism is that our consideration of vehicle list prices does not account for the discounts car buyers usually negotiate, which could partly explain why our respondents underestimated depreciation.

We show that taking typical discounts into account does not change our main conclusions (see SI). Indeed, our approach is conservative because we do not consider other factors that increase the cost of owning a car, such as extra equipment (navigation systems, seat heating and sports seats) or premiums for leasing or financing. These raise car prices by 30–50% on average, and so widen our estimates of ownership-cost misjudgement.

Policy action

It is unlikely that anything will entirely stop people from underestimating the total cost of owning a car. Nevertheless, we think that closing this 'awareness gap' can spur the transition away from conventional cars. How do we do this?

Cars should be labelled with total costs at the point of sale and in registration letters. Such information-provision policies influence consumer purchasing behaviour in a variety of contexts, from buying property to durables such as refrigerators and air conditioning^{17–19}. Many countries, including the United States, Japan and China, already mandate that new cars for sale are labelled with the average future fuel cost of driving them.

Companies that promote alternative forms of transport with lower emissions – such as electric-vehicle dealers, car-sharing or public-transport firms – could boost business by including information on the cost of car ownership in their advertising. To prevent potential conflicts of interest, the information would need to be certified or come from trusted sources, such as scientific institutions or public ministries. But even a general marketing campaign could at least encourage consumers to calculate the cost of driving accurately.

How successful might such interventions be, compared with other options such as a fuel tax or subsidizing public transport? We calculate that fuel prices would need to rise by a massive 1,242% to cut car ownership by the same 37% reduction that we predict as a result

of improved consumer information (see SI). This is because fuel price changes largely target driving itself, rather than the decision of whether to buy a vehicle.

Another widely discussed policy is to subsidize public transport more. This might have less potential than correcting the misjudgement of car ownership costs. There is no evidence available solely from Germany, so we based our extrapolations on evidence for the relationship between prices for public transport and car ownership. This came from a meta study of 83 papers, predominantly from Europe and the United States²⁰. We find that eliminating public-transport fares entirely would decrease car use by only 4.1–6.2%, and could have an even smaller impact on car ownership. This approach would also be burdensome on the public treasury. In Germany, for example, total ticket sales by local transport companies amounted to €13.3 billion in 2019.

We believe that policies on labelling and information would be less costly and less politically fraught than would many other options. The primary resistance to such a policy might

“Fuel price changes largely target driving itself, rather than the decision of whether to buy a vehicle.”

come from conventional car dealers and manufacturers who would be reluctant to see sales fall. But information provision is likely to get strong public support from consumer-protection agencies, for example, which could offset such lobbying. Furthermore, providing ownership cost information could be implemented by revising existing fuel labels, which would reduce institutional obstacles.

Next steps

Future research should focus on total car-ownership costs, rather than on fuel costs alone. To plug knowledge gaps, we have the following recommendations.

We need to know why consumers underestimate the costs of car ownership. This could lead to information targeted to those who make the biggest misjudgements. Campaigns should be tested in the field to guide policymakers who are aiming to promote greener transport.

Our survey should be replicated in other countries to clarify how and where the results apply. This could shed light on what drives the systematic underestimation. Furthermore, future studies should elicit how the underestimation can be reduced, for instance by means of surveys, laboratory experiments and, ideally, field experiments. It would also be useful to investigate the overall impact of car-ownership information on mobility behaviour more broadly, including the use of public transport,

cycling and walking. Our analysis uncovered a need for further empirical evidence on the relationship between the cost of car ownership and the number of cars on the roads.

We see great promise in this research agenda to inform policymakers about cost-effective approaches to reducing emissions from transportation. One of the goals of the European Green Deal, proposed by the European Commission last December, is to accelerate the shift to sustainable and smart mobility. And Horizon Europe, the EU's €100-billion research programme, is currently defining its agenda for research during 2021–27. Both present an invaluable window of opportunity.

The authors

Mark A. Andor is a senior researcher in behavioural, experimental and environmental economics at the RWI — Leibniz Institute for Economic Research, Essen, Germany. **Andreas Gerster** is a postdoctoral researcher in economics at the University of Mannheim, Germany. **Kenneth T. Gillingham** is an associate professor of environmental and energy economics at Yale University, New Haven, Connecticut, USA. **Marco Horvath** is a postdoctoral researcher in environmental and resource economics at the RWI — Leibniz Institute for Economic Research, Essen, Germany. e-mail: kenneth.gillingham@yale.edu

1. International Energy Agency. *Tracking Transport* (IEA, 2019).
2. Oslo Kommune. *The Car-free Livability Programme 2019* (Oslo Kommune, 2019); available at <https://go.nature.com/2vc9ceb>
3. International Energy Agency. *Global EV Outlook 2019* (IEA, 2019).
4. European Environment Agency. *Size of the Vehicle Fleet in Europe* (EEA, 2019).
5. Clark, B., Chatterjee, K. & Melia, S. *Transportation* **43**, 565–599 (2016).
6. von Dauth, W. & Haller, P. [in German] *Berufliches Pendeln zwischen Wohn- und Arbeitsort: Klarer Trend zu längeren Pendeldistanzen* (No. 10/2018) (IAB-Kurzbericht, 2018).
7. Rezvani, Z., Jansson, J. & Bodin, J. *Transp. Res. D* **34**, 122–136 (2015).
8. Krause, R. M., Carley, S. R., Lane, B. W. & Graham, J. D. *Energy Policy* **63**, 433–440 (2013).
9. Dumortier, J. et al. *Transp. Res. A* **72**, 71–86 (2015).
10. Achtnicht, M., Bühler, G. & Hermeling, C. *Transp. Res. D* **17**, 262–269 (2012).
11. Lebeau, K., Van Mierlo, J., Lebeau, P., Mairesse, O. & Macharis, C. *Transp. Res. D* **17**, 592–597 (2012).
12. Bento, A. M., Goulder, L. H., Jacobsen, M. R. & Von Haefen, R. H. *Am. Econ. Rev.* **99**, 667–699 (2009).
13. Allcott, H. *Am. Econ. J. Econ. Policy* **5**, 30–66 (2013).
14. US Environmental Protection Agency. *The Safer Affordable Fuel-Efficient (SAFE) Vehicles Rule for Model Year 2021–2026 Passenger Cars and Light Trucks* (EPA, 2018).
15. Eurostat. *Passenger Cars in the EU* (Eurostat, 2019); available at <https://go.nature.com/3bgchey>
16. Gallagher, K. S. & Muehlegger, E. *J. Environ. Econ. Mgmt* **61**, 1–15 (2011).
17. Davis, L. W. & Metcalf, G. E. *J. Assoc. Environ. Res. Econ.* **3**, 589–625 (2016).
18. Eichholtz, P., Kok, N. & Quigley, J. M. *Am. Econ. Rev.* **100**, 2492–2509 (2010).
19. Newell, R. G. & Siikamäki, J. V. *J. Assoc. Environ. Res. Econ.* **1**, 555–598 (2014).
20. Fearnley, N. et al. *Transp. Res. Procedia* **26**, 62–80 (2017).

Supplementary Information accompanies this article: [see go.nature.com/2kc0796](https://go.nature.com/2kc0796).

Correspondence

Use newfound trust in science wisely

The current COVID-19 pandemic calls for a renewed public trust in science – for better or worse. We urge the global scientific community to seize this opportunity to build on that trust.

Three months into the pandemic, we issued a questionnaire to a panel of 337 US residents who represented a cross-section of the general public. Our aim was to find out how their trust had changed from before the pandemic (data collected in mid-August 2019). Those reporting “a lot of trust” in the federal government remained at an abysmal 1%, whereas “strong trust” in science jumped from 41% to 48%. We found that trust in science was the most important predictor of compliance with public-health recommendations for limiting viral spread.

With great trust comes great responsibility. As we ramp up research to meet the public’s need for solutions, we must be especially careful to communicate transparent information about our capabilities, uncertainties, disagreements or agreements (see S. van der Linden *et al.* *Nature Hum. Behav.* 2, 2–3; 2018).

Competence and warmth are judged by psychologists to be crucial for trustworthiness. Although scientists rate highly on competence, they can sometimes come over as dispassionate (see G. Cardew *Nature* 578, 9; 2020). Now, more than ever, we must show our commitment to humility, honesty and the public good.

Patricia Andrews Fearon, Friedrich M. Götz, David Good
University of Cambridge, UK.
pba21@cam.ac.uk

Better lives call for more than insight

Hetan Shah argues that global problems need social science to help solve them (*Nature* 577, 295; 2020). I contend that he is both right and profoundly wrong.

Developing social inquiry as a social ‘science’ is a blunder that goes all the way back to the eighteenth-century Enlightenment (see go.nature.com/34exatc). To promote human welfare, academia needs to provide practical solutions to problems of suffering, poverty, injustice and avoidable death. It needs to articulate and assess possible solutions in terms of actions, policies, political programmes, philosophies of life and ways of living.

The task of social inquiry and the humanities is to guide people on how to resolve such issues and conflicts in effective, intelligent, humane ways. In connection with the climate crisis, for example, the public needs to know precisely what must be done by governments, businesses, the media, public institutions and individuals to mitigate global warming.

However, social scientists down the decades have fallen short in providing such guidance. In my view, this is because their focus has been on acquiring knowledge about society when it should instead be on promoting social progress towards as good a world as possible.

Nicholas Maxwell University College London, UK.
nicholas.maxwell@ucl.ac.uk

Climate: managing deep uncertainty

In our view, Zeke Hausfather and Glen Peters’s recommendation to assign a single set of best-estimate probabilities to all future emissions scenarios as a means to assess climate-change risks (*Nature* 577, 618–620; 2020) could give decision-makers a false sense of certainty, leading to costly adjustments if the world evolves in unanticipated ways.

The Society for Decision Making Under Deep Uncertainty (www.deepuncertainty.org), to which we belong, offers a better strategy. It relies on methods that focus on the implications of alternative scenarios and the extent to which response tactics are shared across a wide range of scenarios. This helps to manage uncertainties – for example, in sea-level rise after 2050 – by identifying long-term options and short-term, flexible actions that can prepare for a range of future emissions.

Bypassing the need to assign probabilities enables decision-makers to better understand the combination of uncertainties that most affect their choices, thereby reducing locked-in choices and decision delays that can arise when using a single scenario.

Judy Lawrence Victoria University of Wellington, New Zealand.
judy.lawrence@vuw.ac.nz

Marjolijn Haasnoot Deltares and University of Utrecht, Utrecht, the Netherlands.

Robert Lempert RAND Corporation, Santa Monica, California, USA.

Climate: why use 2100 for timeline?

Zeke Hausfather and Glen Peters’s discussion of future climate scenarios focuses on what we might expect by 2100 (*Nature* 577, 618–620; 2020). But why 2100? This inordinate focus on the century’s end, largely derived from Intergovernmental Panel on Climate Change scenarios, has coloured much of the literature for years and now saturates the public debate.

Take, for instance, the authors’ tags for warming above pre-industrial levels: 1.5 °C, “mitigation required to reach Paris goals”; 2.5 °C, “modest mitigation”; 3 °C, “weak mitigation (likely)”; 4 °C, “average no policy (unlikely)”; 5 °C, “worst-case no policy (highly unlikely)”. Peak warming will post-date peak emissions and, depending on feedbacks, the planet will still be warming in 2100 – even in some of the “likely” pathways and certainly in the “unlikely” and “highly unlikely” ones.

Let’s move the discussion to peak impact and a full-recovery timescale, especially when considering policy.

Paul N. Pearson Cardiff University, UK.
pearsonp@cardiff.ac.uk

HOW TO SUBMIT

Correspondence may be submitted to correspondence@nature.com after consulting the author guidelines and section policies at go.nature.com/cmchno.

News & views

Volcanology

When it rains, lava pours

Michael Manga

Early 2018 saw unusually heavy rainfall in Hawaii. Modelling now suggests that groundwater pressure increased owing to rainfall: this might have triggered changes in the eruption of the island's Kilauea volcano. **See p.491**

The most recent eruption of Kilauea volcano on the island of Hawaii began in 1983. For 35 years, most of its magma emerged from a set of fissures in the volcano called the upper east rift zone. But on 3 May 2018, Kilauea's lower east rift zone opened up, giving way to a massive outpouring of lava that devastated the southeastern part of the island¹ (Fig. 1). An important question is why this change occurred in May 2018, rather than earlier or later in the course of the eruption. On page 491, Farquharson and Amelung² propose that record-breaking levels of rainfall in early 2018 increased groundwater pressures which, in turn, made it easier for rock to break and hence magma to rise to the surface at new locations.

The creation of a pathway that brings magma to Earth's surface begins with the mechanical failure of rocks. This failure can occur in two ways: new cracks can open, or existing faults can slip. Both processes can be promoted by pressure changes in groundwater. For the former, increases in fluid pressure decrease the amount of stress needed to open new cracks. For the latter, faults can slip when the stresses acting parallel to the fault (shear stresses) overcome those perpendicular to the fault (normal stresses). These normal stresses act to clamp the fault shut. Increasing fluid pressure in rocks lowers normal stresses without changing shear stresses, thus promoting fault failure.

Heavy rainfall increases water levels underground and thus pressure in groundwater. The volcanic rocks in Hawaii are very permeable, which allows water to infiltrate and pressure changes to propagate to a depth of several kilometres, close to where magma is stored. Fluid-pressure changes take time to propagate from the surface to those depths. Thus, downward migration of rock failure over time, along with a time lag between the accumulation of water at the surface and failure at

depth³, would be key indicators that rainfall was the cause of rock failure at Kilauea.

Farquharson and Amelung modelled pressure changes at Kilauea caused by rainfall in the months leading up to the eruption on 3 May 2018. Their model showed an increase in pressure of tens to hundreds of pascals at depths of several kilometres. On the basis of these changes, along with four sets of observations indicating that eruptions at Kilauea are associated with patterns of substantial rainfall, the authors propose that heavy rainfall promoted the rock failure that enabled magma to flow into the lower east rift zone.

Is their hypothesis plausible? The pressure changes computed by their models are small – smaller than stresses from tides. However, if rocks are already close to breaking, such changes might be sufficient to initiate failure. The 2018 eruption was accompanied by

a magnitude-6.9 earthquake, and examples of earthquakes caused by pressure changes on this scale are abundant⁴. For example, the widespread increase in earthquake frequency in the central and eastern United States in the past decade results from wastewater injection into permeable rocks that increases water pressure and changes stresses⁵.

The geological record also confirms that changes in stresses at Earth's surface can modulate volcanic activity. Onland, volcanism is promoted by the retreat of glaciers⁶. Sea-level changes between glacial and interglacial periods can modulate eruption rates at mid-ocean ridges⁷. Stresses from large earthquakes increase the probability of volcanic eruptions⁸ and can change activity at volcanoes that are already active⁹.

Although it is well established that changes in water pressure promote earthquakes, they are not necessarily a direct cause of magma eruption. To begin moving through Earth's crust, magma must create large enough stresses in the surrounding rocks to open a pathway. Earthquakes triggered in the crust around that stored magma, however, can actually relieve stress – as such, they might make it more difficult for magma to erupt¹⁰.

Ultimately, whether fault failure from water-pressure changes can occur close to stored magma, as hypothesized by Farquharson and Amelung, remains uncertain. The first magma to erupt from the lower east rift zone in 2018 was old, perhaps left over from an earlier, 1955 eruption¹¹, implying that the rift zone was already hot. As a result, groundwater



Figure 1 | Lava from the lower east rift zone of Kilauea volcano. Farquharson and Amelung² propose that exceptionally heavy rainfall led to the eruption of magma from this part of the volcano in 2018.

in the rift zone might have been vapour at shallow depths¹², and at greater depths it could have been a supercritical fluid (a substance that is not in a distinct liquid or gas phase, but has properties of both). The high compressibility of both vapours and supercritical fluids would dampen the magnitude of pressure changes in the authors' model, making failure less probable.

How, then, can we test the hypothesis that rainfall initiated the lower east rift zone eruption? Unfortunately, subsurface pressure measurements – and hydrogeological data more generally – are rarely part of volcano monitoring. Instead, as with many geoscience and Earth-history questions, we have to look back in time using the geological and historical record of eruptions. In support of their hypothesis, Farquharson and Amelung analysed all reported eruptions at Kilauea since 1790, and showed that the volcano tends to erupt at the wettest time of year.

Should we increase alert levels at volcanoes after heavy rainfall? We could ask the same question about other stress changes, such as those from regional earthquakes. This is an open question. These stress changes are small, and hence, if anything, modulate the exact timing of the surface eruption. At Kilauea, there were other sources of stress – in fact, a change in eruption behaviour had been anticipated on the basis of ground-deformation measurements and inferred magma movement. The Hawaiian Volcano Observatory issued a warning on 17 April that a new vent might open¹.

The possibility that external processes initiate volcanic eruptions is a reminder that volcanoes are part of a dynamic Earth system. Volcanic eruptions influence all surface environments, including climate and weather¹³. Changes in those surface environments, such as heavy rainfall, might also influence eruptions. We are only just beginning to understand these interactions.

Michael Manga is in the Department of Earth and Planetary Science, University of California, Berkeley, Berkeley, California 94720-4767, USA. e-mail: manga@seismo.berkeley.edu

1. Neal, C. A. *et al.* *Science* **363**, 367–374 (2019).
2. Farquharson, J. I. & Amelung, F. *Nature* **580**, 491–495 (2020).
3. Montgomery-Brown, E. K., Shelly, D. R. & Hsieh, P. A. *Geophys. Res. Lett.* **46**, 3698–3705 (2019).
4. Foulger, G. R., Wilson, M. P., Gluyas, J. G., Julian, B. R. & Davies, R. J. *Earth Sci. Rev.* **178**, 438–514 (2018).
5. Keranen, K. M., Weingarten, M., Abers, G. A., Bekins, B. A. & Ge, S. *Science* **345**, 448–451 (2014).
6. Jellinek, A. M., Manga, M. & Saar, M. O. *J. Geophys. Res. Solid Earth* **109**, B09206 (2004).
7. Boulahanis, B. *et al.* *Earth Planet. Sci. Lett.* **535**, 116121 (2020).
8. Linde, A. T. & Sacks, I. S. *Nature* **395**, 888–890 (1998).
9. Avouris, D. M., Carn, S. A. & Waite, G. P. *Geology* **45**, 715–718 (2017).

10. Gudmundsson, A. *Volcanotectonics: Understanding the Structure, Deformation and Dynamics of Volcanoes* (Cambridge Univ. Press, in the press).
11. Gansecki, C. *et al.* *Science* **366**, eaaz0147 (2019).
12. Hsieh, P. A. & Ingebritsen, S. E. *J. Geophys. Res. Solid Earth* **124**, 1498–1506 (2019).

13. National Academies of Sciences, Engineering, and Medicine. *Volcanic Eruptions and Their Repose, Unrest, Precursors, and Timing* (National Academies, 2017).

Condensed-matter physics

Permanent electric control of spin current

Stefano Gariglio

The development of low-power methods for controlling a property of electrons known as spin could help to maintain the historic rates of progress that are occurring in computational power. Just such a method has now been reported. **See p.483**

A promising technology for the next generation of computers is spintronics, a type of electronics that depends on the spin – the intrinsic angular momentum – of electrons, rather than their charge. However, available methods for controlling spin require electric currents that are too large for practical applications. On page 483, Noël *et al.*¹ report an approach that allows low-power spin control using an electric field.

The exponential progress in increasing computational power over the past 50 years has been largely driven by the relentless miniaturization of the field-effect transistor², the basic component of silicon chips. This consistent downscaling was anticipated³ in 1965 by electronic engineer Gordon Moore, and has led to the staggering 2 billion transistors that are now typically found in the processors of modern personal computers. The semiconductor industry has come up with a road map outlining the technological developments in computer materials, devices and systems that will be needed to maintain these historic rates of increase in computational power (<https://irds.ieee.org>).

A growing section of the road map addresses a pressing problem for the field: transistors based on currently used technology cannot be scaled down much further, because the physical limits of miniaturization will soon be reached. There are no known solutions for several of the technical and materials issues associated with this problem. Materials scientists, physicists and engineers are therefore investigating an array of potential new working principles for computer technology. The development of new approaches also allows other goals to be targeted, such as lowering energy consumption, or incorporating multiple functionalities into components to speed up data processing.

One way of reducing power consumption

would be to eliminate the need for a continuous power supply to maintain the logic state (ON or OFF) of transistors. This can be achieved using ferroic materials (such as ferroelectric compounds, which have a permanent electric polarization) or piezoelectric mechanical devices, which require power to switch between the logic states, but not to retain those states⁴. Spintronics technology has also seen a surge of interest, because this approach is expected to reduce electrical dissipation⁵ – wasteful loss of electrical power as heat. Combinations of ferroic approaches with spintronics⁶ could be particularly effective in the race to develop more-efficient computing technology.

However, many of these approaches will require new materials – for example, the semiconductors used in conventional electronic devices do not have ferroic properties. A family of compounds known as complex oxides are of particular interest, because they host permanent electric and magnetic dipoles, thereby opening the door to applications that require permanent states. Although complex oxides are not as good as semiconductors for use in classical transistors because they produce more electrical dissipation, they have remarkable properties for spintronics⁷.

Interesting electronic phases have been observed to form at the interfaces between two complex oxides. Noël and co-workers focus on a phase called an electron gas: an ultrathin (a few nanometres thick) layer of conducting electrons that forms at the surface of strontium titanate (STO) that has been covered by a layer of aluminium.

STO probably provides the best illustration of the complexity of the electrical properties of transition-metal oxides. In its pure form, it is a dielectric material (an electrical insulator) that has a tendency to become ferroelectric at temperatures below 4 kelvin, but fails to do so

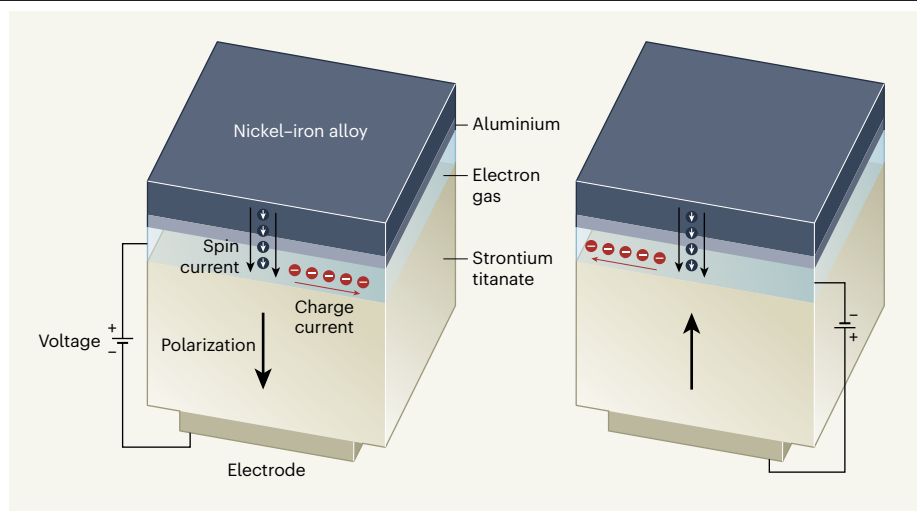


Figure 1 | A ‘ferroelectric-like’ spin–orbit transistor. A type of electronics known as spintronics involves controlling the spin (the intrinsic angular momentum) of electrons. Noël *et al.*¹ injected a spin current (arrows in circles indicate electron spins) from a magnetic nickel–iron alloy into strontium titanate (STO). The STO was covered by a thin layer of aluminium, which induces the formation of an electron gas (a layer of highly mobile electrons) at the STO surface. The electrons in the gas exhibit spin–orbit coupling – their spins couple to their momenta. This effect converts the spin current into a conventional charge current (red circles indicate electron charges). The authors applied a voltage across the insulating STO beneath the electron gas to change the sign of the spin–orbit coupling, and hence the direction of the charge current. The insulating STO exhibits surprising ‘ferroelectric-like’ behaviour: it has an overall electrical polarization whose direction (large arrow) depends on the applied voltage. The polarization remains in the absence of an electric field; this allows permanent control of the spin–orbit coupling and thereby of the direction of the charge current.

because of quantum fluctuations⁸. However, tweaks to the chemistry of STO (such as the replacement of some of the strontium atoms by calcium atoms) can push the compound over the edge to become truly ferroelectric⁹. And the replacement of some of the strontium atoms by lanthanum atoms increases the number of electrons in STO through a process known as electron doping, and turns the material into a metallic conductor, and even into a superconductor¹⁰.

One consequence of the electrons at the aluminium/STO interface becoming confined in a gas is that their spin is coupled to their momentum, a phenomenon called spin–orbit interaction¹¹. Workers from the same research group as Noël *et al.* have previously demonstrated¹² such spin–momentum entanglement: when they injected a spin-polarized current (a flow of spins that are oriented in one direction) into the electron gas, they observed a conventional electric current (a charge current) whose direction depends on the spin orientation and on the spin–orbit coupling. This is the result of the spin polarization being converted into electron motion by the spin–orbit interaction.

Noël *et al.* now report surprising observations of the STO electron-gas system that adds to this complex behaviour. When the authors applied an electric field to the STO to control spin–orbit coupling in the gas, they observed a hysteresis effect – the direction of the charge current produced in the gas ‘remembers’ the

polarity of the applied electric field, even after the field is removed (Fig. 1). Moreover, when they characterized the properties of the insulating STO beneath the gas, they observed features commonly attributed to ferroelectric compounds: when the polarity of the voltage applied across the STO is reversed, a spike of charge current is produced. Such a phenomenon is commonly associated with the reversal of electric dipoles in ferroelectric materials, and is at the core of the definition of electrical polarization in the modern theory of ferroelectricity¹³.

The authors’ system has potential applications for spintronics, because it acts as a spin detector, analogous to optical polarizers that transmit light polarized only along a particular direction. Moreover, when the polarity of the applied voltage is inverted, the selectivity of the spin filter changes so that electrons with ‘spin up’ polarization move right, instead of left. Crucially, this selectivity remains in the absence of an applied voltage. This minimizes power consumption and opens up applications for memory storage.

Ferroelectric-like behaviour has been observed previously in STO (see refs 14–16, for example). However, a peculiarity of the effect observed by Noël and colleagues is that it occurs only when the applied electric field exceeds a critical value. This raises concerns: true ferroelectric materials don’t show polarization only at high fields; it is an intrinsic state that occurs even in the absence of an

external electric field. Noël and colleagues’ data indicate that something similar to a ‘relaxor’ state occurs in their system at low temperatures, in which a fraction of the STO consists of nanometre-scale domains that have an electrical polarization, and move or reorient in an applied electric field. By contrast, all of the material in a true ferroelectric compound is polarized.

One can speculate that the movement of polar walls¹⁷ – boundaries that form between two STO domains that have different crystallographic orientations¹⁸ – produces the spikes of current observed by Noël and co-workers. But other microscopic mechanisms might be at play, given the richness of STO’s electronic behaviour; point defects produced in STO during the fabrication of the authors’ device could also have a role. Research into the domain walls in STO is currently booming, and will probably find an explanation for the observed behaviour. In the meantime, the authors’ demonstration of a permanent switch of spin–orbit coupling at a complex-oxide interface shows the potential of this class of material to compete in the race for more-efficient computing.

Stefano Gariglio is in the Department of Quantum Matter Physics, University of Geneva, Geneva CH-1211, Switzerland.
e-mail: stefano.gariglio@unige.ch

1. Noël, P. *et al.* *Nature* **580**, 483–486 (2020).
2. *Nature* **479**, 309 (2011).
3. Moore, G. E. *Electronics* **38**, 114–117 (1965).
4. Abele, N. *et al.* *IEEE Int. Electron Devices Meet., 2005. IEDM Tech. Digest* 479–481 (2005).
5. Joshi, V. K. *Eng. Sci. Technol.* **19**, 1503–1513 (2016).
6. Manipatruni, S., Nikonov, D. E. & Young, I. A. *Nature Phys.* **14**, 338–343 (2018).
7. Förg, B., Richter, C. & Mannhart, J. *Appl. Phys. Lett.* **100**, 053506 (2012).
8. Müller, K. A. & Burkard, H. *Phys. Rev. B* **19**, 3593–3602 (1979).
9. Bianchi, U., Kleemann, W. & Bednorz, J. G. *J. Phys. Condens. Matter* **6**, 1229 (1994).
10. Schooley, J. F., Hosler, W. R. & Cohen, M. L. *Phys. Rev. Lett.* **12**, 474–475 (1964).
11. Caviglia, A. D. *et al.* *Phys. Rev. Lett.* **104**, 126803 (2010).
12. Lesne, E. *et al.* *Nature Mater.* **15**, 1261–1266 (2016).
13. Resta, R. *Ferroelectrics* **136**, 51–55 (1992).
14. Sidoruk, J. *et al.* *Ferroelectrics* **505**, 200–209 (2016).
15. Manaka, H., Nozaki, H. & Miura, Y. *J. Phys. Soc. Jpn* **86**, 114702 (2017).
16. Hemberger, J., Lunkenheimer, P., Viana, R., Böhmer, R. & Loidl, A. *Phys. Rev. B* **52**, 13159–13162 (1995).
17. Schiaffino, A. & Stengel, M. F. *Phys. Rev. Lett.* **119**, 137601 (2017).
18. Honig, M. *et al.* *Nature Mater.* **12**, 1112–1118 (2013).

Ecology

The pace of biodiversity change in a warming world

Jennifer M. Sunday

The timing of disruptions to biodiversity associated with global warming is a key, but little-explored, dimension of change. Will losses in biodiversity occur all at once, or be spread out over time? **See p.496**

Projections of the effects of climate change on multiple species are often made by estimating the change predicted for a single future time point; for example, by asking how the geographical distributions of multiple species will differ in 2100 from those today¹. However, this approach does not capture the pace, timing or possible synchrony of biodiversity changes across time. Acute synchronous impacts can potentially be more damaging to a system than those spread over time, in terms of both human adaptation to biodiversity losses and ecosystem resilience. On page 496, Trisos *et al.*² report an approach for predicting how climate change will affect future biodiversity patterns.

The authors estimated the timing and synchrony of climate impacts on organisms globally by asking when species in a given region will be exposed to temperatures outside their normal global experience (by considering projected future temperatures due to climate change). They did this by compiling geographical-range maps for approximately 30,000 species, including birds, mammals, reptiles, amphibians, fishes, marine invertebrates, corals and seagrasses, and using temperature-projection models to identify the warmest average annual temperature experienced between 1850 and 2005 by each species within its range. Dividing Earth into grid cells of 100 square kilometres and using predicted climate information, the authors determined when each species would experience annual average temperatures above its historical annual average, encountered anywhere in its range, for an extended period. The result provides an estimate of when a species will be exposed to unprecedentedly high temperatures.

Trisos and co-workers' approach builds on 'time of emergence', a concept used when analysing climate change. Time of emergence describes the time at which a climate variable, such as temperature, emerges beyond the historical values of variation observed for a particular location – in other words, when the

average value of the measurement of interest becomes more extreme than the previously encountered natural variability. Trisos *et al.* offer innovation in applying this concept to the realm of biodiversity. First, rather than considering the variation experienced at just one location, they considered the full breadth of variation experienced across each species' geographical range, defining an organism as being 'exposed' in a specific grid cell only after it has experienced temperatures above its range-wide maximum (and with annual temperatures remaining above this value for a minimum of five years). Second, because the authors considered multiple species in an assemblage (the group of species present

in a given grid cell), it was possible to assess the relative timing of exposure in a graphical format that the authors call a horizon profile (Fig. 1). This enables the synchrony in the timing of exposure events for the species in a region to be quantified and easily visualized.

The authors' results predict that the greatest levels of exposure will occur at latitudes nearer the Equator, and, most notably, that there will be high synchrony in the timing of exposure between species in the same grid cell, for grid cells both on land and in the ocean. Trisos *et al.* find that most species in a given cell will usually become exposed to unprecedentedly high temperatures within the same decade. If this exposure results in local extinction, it suggests the following disturbing scenario. We might initially see a small trickle of species being lost from an assemblage, but this will be followed by an abrupt loss of most species in the assemblage within the same decade.

What mechanism might explain this predicted pattern? The abruptness in exposures predicted by Trisos and colleagues is not due to any particular abruptness in the timing of climate change itself – although similar predictions of abrupt ecological change have been based on the additive effects of gradual climate change with abrupt natural climate variability, including weather³. Instead, it seems to be attributable to the similarity of the thermal niches occupied by the species in each grid cell. Trisos and

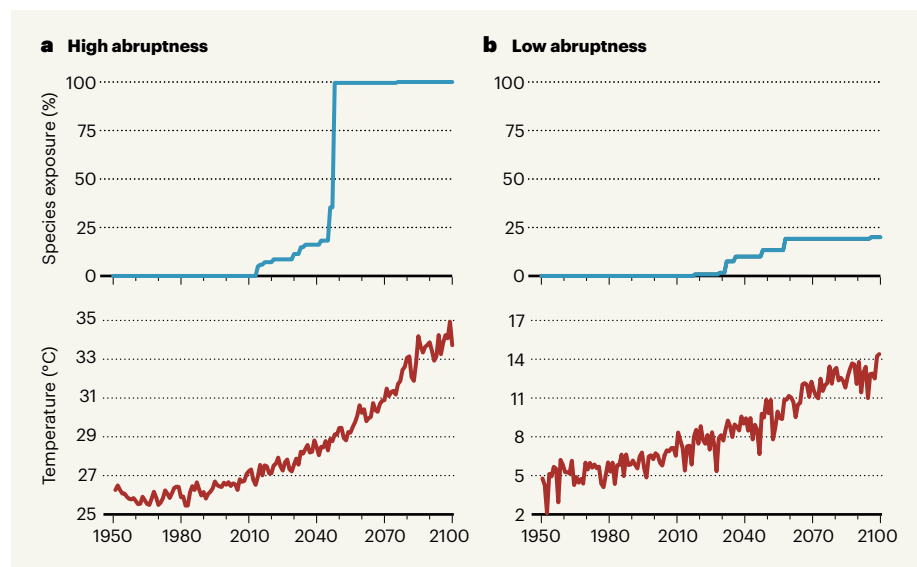


Figure 1 | Exposure of species to unprecedented temperatures owing to climate change. Trisos *et al.*² report a global analysis describing when species (as assessed in grids of 100 square kilometres) are predicted to encounter, for more than five years, higher annual average temperatures than they have previously experienced anywhere in their geographical range. This state is called species exposure, and the authors assessed more than 30,000 terrestrial and marine species. **a**, For the majority of locations, the timing of when most species in a grid are exposed occurs highly abruptly, as in this example from the Amazon basin. Temperatures are from projection models, and the predicted future temperatures due to climate change are from the RCP8.5 model, which depicts a scenario of high greenhouse-gas emissions⁷. **b**, By contrast, in the Gobi Desert (located in northern China and southern Mongolia), species exposure occurs with low abruptness. (Graphs based on Fig. 1 of ref. 2.)

colleagues find that more than half of the species in a given cell (and almost 90% in most marine assemblages) tend to have geographical ranges that encompass similarly warm temperatures, such that they would all face exposure at around the same time.

Such a striking pattern of shared thermal niches within assemblages has been observed before, in a global analysis of marine fishes and invertebrates⁴. In that study, species' thermal niches were found not to change gradually with latitude, but instead to have distinct transition points, indicating that species belong to what are termed thermal guilds⁴. These shared thermal niches could be due to physical boundaries or ecological interactions that restrict the ranges – and temperatures experienced – of multiple species similarly. Or this phenomenon might be the result of a low rate of evolution in the range of temperatures across which the species can fundamentally persist, leading to the maintenance of thermal guilds.

When does this abrupt exposure happen? It is predicted that it will occur at different times for grid cells around the world, from some predicted to be occurring already in the ocean, to others occurring towards the end of the projected time range, in 2100. That the timing is different across grid cells is a good thing, because at least all of the assemblages aren't predicted to experience abrupt losses at the same time. But, notably, the timing of exposure does not correlate with the timing of climate-change emergence in temperature, suggesting that the latter metric might be a poor predictor of major biodiversity change within a given grid cell.

Trying to project the timing of biodiversity shifts is a noble objective that will surely help us to develop management systems and anticipate crises. Although Trisos *et al.* provide an initial approach that offers useful insights, further studies should attempt to validate and qualify these predictions. For example, Trisos and colleagues used temperatures outside species' current thermal niches to define climate exposure, but we don't know what will really occur when species experience such temperatures – many can certainly tolerate temperatures beyond those found in their current ranges^{5,6}. The timing of exposure to truly limiting environments might turn out to be more diverse across species than currently predicted by Trisos *et al.* if variation in species' fundamental climatic niches (the range of temperatures and other climate variables across which an organism can survive) is considered. It will also be useful to consider the flip side of the range-shift issue: the timing and abruptness with which new species enter an assemblage as a result of range extensions arising from climate change.

Most crucially, as climate change progresses, we should be able to test and refine

projections such as these using real-time observations. Where are biodiversity changes already occurring abruptly? The need for systematic global biodiversity monitoring has never been stronger.

Jennifer M. Sunday is in the Department of Biology, McGill University, Montreal, Quebec H3A 1B1, Canada.
e-mail: jennifer.sunday@mcgill.ca

Evolution

Can't see the wood for the trees

Mark Pagel

Evolutionary-tree diagrams, which show the branching relationships between species, are widely used to estimate the rates at which new species arise and existing ones become extinct. New work casts doubt on this approach. **See p.502**

Scientists often want to make inferences about what the biological past was like, and how that past gave rise to the present, because doing so allows them to understand the processes that drive evolution. But on page 502, Louca and Pennell¹ challenge a major aspect of that enterprise.

Specifically, their work regards the issue of estimating past rates of speciation and extinction, which are, respectively, the rates at which new species arise and existing species go extinct. These rates determine the number of contemporary species of various forms. There are, for instance, around 6,600 species

“Assumptions are being made about the things that we would like to estimate.”

of songbird (passerines), which constitute more than half of all existing bird species, and we might therefore be tempted to say that songbirds have a high rate of speciation in comparison with that of other birds. But it's also possible to speculate that they have a low extinction rate. Louca and Pennell show that the uncertainty is even worse than this: not only can we not estimate these two rates, but also there is an infinite number of different sets of these two parameters that are equally good at describing any particular outcome, such as the number of species of contemporary songbird.

Because fossils are scarce or non-existent

1. Warren, R., Price, J., Graham, E., Forstenhaeusler, N. & VanDerWal, J. *Science* **360**, 791–795 (2018).
2. Trisos, C. H., Merow, C. & Pigot, A. L. *Nature* **580**, 496–501 (2020).
3. Harris, R. M. B. *et al.* *Nature Clim. Change* **8**, 579–587 (2018).
4. Stuart-Smith, R. D., Edgar, G. J., Barrett, N. S., Kininmonth, S. J. & Bates, A. E. *Nature* **528**, 88–92 (2015).
5. Sunday, J. M., Bates, A. E. & Dulvy, N. K. *Nature Clim. Change* **2**, 686–690 (2012).
6. Early, R. & Sax, D. F. *Glob. Ecol. Biogeogr.* **23**, 1356–1365 (2014).
7. van Vuuren, D. P. *et al.* *Clim. Change* **109**, 5–31 (2011).

This article was published online on 8 April 2020.

for the vast majority of species, evolutionary scientists instead estimate speciation and extinction rates from phylogenies – tree diagrams that describe the patterns of descent among a group of contemporary species (Fig. 1a,b). For any such phylogeny, it is easy to construct what is termed a lineage-through-time plot; this records the cumulative number of lineages up to that point in time on the tree that will eventually leave one or more living descendent species (Fig. 1c). The slope of the curve fitted to such a plot, often denoted by λ , is the net speciation rate. This is equal to the difference between the rate of speciation, termed b (or birth), and the rate of extinction, termed d (or death). It is described by the equation $\lambda = b - d$.

However, it is known that a difficulty arises in estimating b and d , because if all that is available is the number of species that have survived to the present, such as our 6,600 songbirds, any pair of b and d that returns the same value of λ will produce an identical lineage-through-time curve, and there is an infinite number of these pairs. In fact, it turns out that for the simple case of estimating $b - d$, such as described here, a feature of the shape of the lineage-through-time curve can be exploited to estimate the rate of extinction, and then the rate of speciation can be found by subtraction². But to do so requires making the assumption that both of these rates are constant throughout the entire time span of the tree, when instead they almost certainly vary between the different branches (lineages) of the phylogeny, and through time.

This is where Louca and Pennell step in,

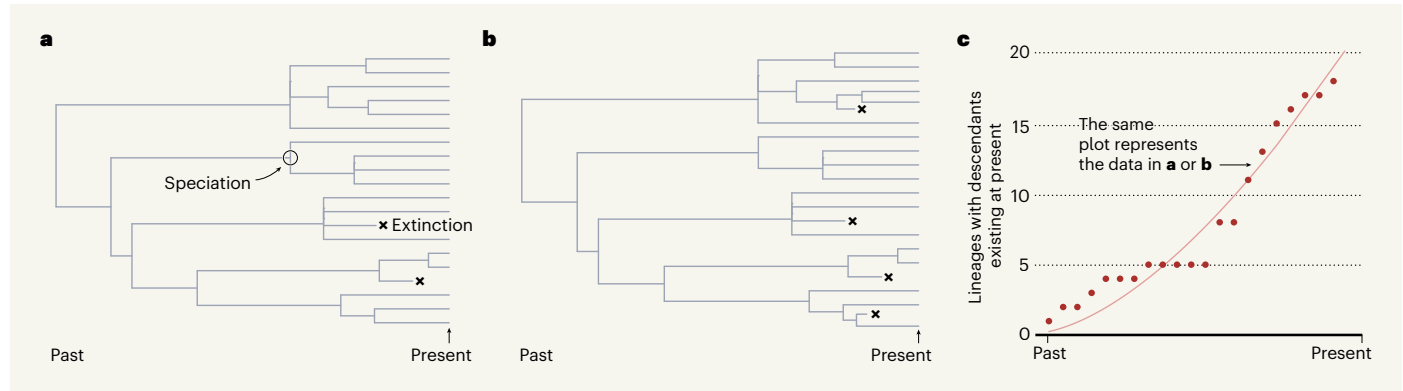


Figure 1 | Assessing evolutionary histories. Louca and Pennell¹ raise questions about a standard approach to estimating past rates of species formation (speciation) and extinction that uses data from a lineage-through-time plot. The number of species in the present depends on how speciation and extinction rates varied over time in the past. Using mathematical modelling, the authors reveal that an infinite number of pairs of speciation and extinction rates could give rise to any given outcome, and it is thus unclear how to determine the correct rates. **a, b**, Examples of known extinctions are rare, and are shown in these hypothetical tree diagrams only

to illustrate how different rates of extinction (and different speciation rates) can yield the same lineage-through-time plot. **c**, Information taken from a tree diagram can be represented in a lineage-through-time plot as shown. Red dots indicate the number of lineages at a given time that gave rise to lineages existing in the present. The slope of the curve equals the speciation rate minus the extinction rate. This plot is valid for both trees even though they have different speciation and extinction rates. This underscores the authors' demonstration that many different data inputs can give identical lineage-through-time plots.

because the novelty and mathematical sophistication of their work lie in showing that we cannot estimate these 'time-varying' speciation and extinction rates. The authors invoke earlier work³ that defines the existence of a tree's 'deterministic' lineage-through-time curve: this is a set of differential equations (equations describing rates of change) that fully determine the number of lineages in a tree at any given time. Louca and Pennell's key result is then to show that there is an infinite number of alternative sets of time-varying speciation–extinction rates that yield the same number of lineages at any given time as does the deterministic lineage-through-time curve. They further show that the most probable estimates of the two rates (calculated by maximum-likelihood methods) do not necessarily identify the correct underlying model – as demonstrated by an analysis of hypothetical cases for which the true time-varying speciation–extinction rates are known.

Even worse for those who want to use the rates of speciation and extinction to study evolution, the possible alternative scenarios of time-varying speciation and extinction rates that are consistent with the deterministic lineage-through-time model often differ qualitatively. For example, the authors show that a phylogeny of approximately 80,000 species of seed plant is equally well described by speciation and extinction rates that both gradually increase through time or that both gradually decrease through time. Other scenarios, including rates that vary wildly with time, provide equally good descriptions of the numbers of lineages through time as derived from the deterministic lineage-through-time model.

Louca and Pennell's conclusions will be

dispiriting to evolutionary scientists who are looking for a link between past levels of speciation and extinction and historical climate change or other environmental events, or who want to test ideas about what features of a species – such as diet, mating system or the length of a generation – might be used to predict speciation and extinction rates⁴. The limitations that Louca and Pennell have identified for estimating speciation and extinction rates do not go away as the size of the phylogenetic tree increases. Nor do other common features of trees provide much help: for example, if a group of species has never suffered any extinctions, estimating their speciation rate would be straightforward. But this is rare, and unlikely to be known in advance. Having abundant fossils could help, because they provide evidence needed to estimate extinction rates; however, fossils are seldom abundant. We can make assumptions about how speciation and extinction might vary with each other, through time, or with the number of species, but these assumptions are being made about the things that we would like to estimate.

Amid this epistemological carnage regarding what we can possibly know, the authors helpfully offer some consolation by showing that it is possible to estimate a parameter they call the pulled speciation rate, or λ_p . This measures the rate of change (the slope of the curve) of the deterministic model of the lineage-through-time plot. The pulled speciation rate can be compared between lineages, or at different times, and might be useful for understanding the processes that gave rise to the species that are alive today, even if not necessarily providing information about those species that didn't make it.

And this aspect – the ones that became extinct – is the deeper lesson of Louca and

Pennell's work. Without fossils, all evolutionary scientists, whether studying speciation and extinction or attempting to reconstruct the features of distant ancestors, need to be aware that the evolutionary processes they identify are those that operated in the species that would survive and eventually leave descendants in the present. We can't be sure what was going on in those that went extinct. It is the evolutionary version of the observation that history is written by the victors. The supreme irony of this predicament is that Charles Darwin's idea about the survival of the fittest, the story that we want to understand, by its very nature renders elusive some of the key components needed to study it.

Mark Pagel is at the School of Biological Sciences, University of Reading, Reading RG6 6UR, UK.
e-mail: m.pagel@reading.ac.uk

1. Louca, S. & Pennell, M. W. *Nature* **580**, 502–505 (2020).
2. Nee, S., Holmes, E. C., May, R. M. & Harvey, P. H. *Phil. Trans. R. Soc. Lond. B* **344**, 77–82 (1994).
3. Kubo, T. & Iwasa, Y. *Evolution* **49**, 694–704 (1995).
4. Morlon, H. *Ecol. Lett.* **17**, 508–525 (2014).

This article was published online on 15 April 2020.

The wide-binary origin of (2014) MU₆₉-like Kuiper belt contact binaries

<https://doi.org/10.1038/s41586-020-2194-z>

Evgeni Grishin^{1✉}, Uri Malamud¹, Hagai B. Perets¹, Oliver Wandel² & Christoph M. Schäfer²

Received: 19 February 2020

Accepted: 27 February 2020

Published online: 22 April 2020

 Check for updates

Following its flyby and first imaging of the Pluto–Charon binary, the New Horizons spacecraft visited the Kuiper belt object (KBO) 2014 MU₆₉ (also known as (486958) Arrokoth). The imaging showed MU₆₉ to be a contact binary that rotates at a low spin period (15.92 hours), is made of two individual lobes connected by a narrow neck and has a high obliquity (about 98 degrees)¹, properties that are similar to those of other KBO contact binaries inferred through photometric observations². However, all scenarios suggested so far for the origins of such configurations^{3–5} have failed to reproduce these properties and their probable frequent occurrence in the Kuiper belt. Here we show that semi-secular perturbations^{6,7} operating on only ultrawide KBO binaries close to their stability limit can robustly lead to gentle, slow binary mergers at arbitrarily high obliquities but low rotational velocities, reproducing the characteristics of MU₆₉ and other similar oblique contact binaries. Using *N*-body simulations, we find that approximately 15 per cent of all ultrawide binaries with a cosine-uniform inclination distribution^{5,9} are likely to merge through this process. Moreover, we find that such mergers are sufficiently gentle to deform the shape of the KBO only slightly. The semi-secular contact binary formation channel not only explains the observed properties of MU₆₉, but may also apply to other Kuiper belt or asteroid belt binaries and in the Solar System and extra-solar moon systems.

The discovery of the bilobate shape of MU₆₉ and its peculiar configuration provided new clues and opened avenues of exploration into the physical processes that sculpt the Solar System. Here we describe an evolutionary channel for the formation of MU₆₉ from an initially wide binary. We consider the initial binary to be a member of a hierarchical triple together with the Sun. Owing to secular evolution induced by the Sun, the inner orbit may experience changes in its eccentricity (*e*) and mutual inclination (*i*) on secular timescales much longer than the orbital period, known as Lidov–Kozai (LK) oscillations, which can be modelled using a secular orbit-averaging approach^{10,11}. Large LK oscillations take place when the mutual inclination is large ($40^\circ \leq i \leq 140^\circ$). The highest eccentricities are attained as the binary evolves to the lowest inclinations and vice versa¹².

If the eccentricity of the binary exceeds a threshold e_{coll} , the small pericentre allows binary collisions. Thus, LK evolution could lead to coalescence of individual Kuiper belt binary (KBB) members into a single, probably irregularly shaped, KBO⁵. However, because the closest approach occurs concurrently with the lowest inclinations, collisions mostly occur near $i \approx 40^\circ$ and $i \approx 140^\circ$ (ref. ¹³). Moreover, tidal effects and the non-spherical structure of KBB components quench LK evolution, which makes collision possible only in a small part of the parameter space^{5,14}. The standard LK mechanism is therefore disfavoured for the origin of the highly oblique MU₆₉, but can explain the origin of highly eccentric KBBs such as WW31 and 2001 QW322^{5,15,16}.

For larger ratios of the inner period to outer period, secular averaging breaks down and the evolution becomes semi-secular. The orbit of

the inner binary now evolves considerably on timescales of the outer orbit, and short-term fluctuations arise, making the LK evolution more complex^{6,7,17}. The maximal eccentricity can be calculated analytically, including in domains where it is unconstrained⁷ and the evolution is non-secular. Figure 1a shows the analytical two-dimensional parameter space for allowed and forbidden domains for collisions in terms of the initial inclination $\cos i_0$. The initial separation of the inner binary is normalized to the Hill radius, $r_{\text{H}} = a_{\text{out}}(m_{\text{in}}/3M_{\odot})^{1/3}$, where a_{out} is the outer semi-major axis, M_{\odot} is the mass of the Sun and m_{in} is the mass of the inner binary. For an inner semi-major axis *a*, the (dimensionless) separation $\alpha \equiv a/r_{\text{H}}$ cannot exceed the Hill stability limit for highly inclined orbits⁸, $\alpha_{\text{H}} = 0.4$. We use the outer orbit parameters of MU₆₉: $a_{\text{out}} = 44.581$ AU and eccentricity $e_{\text{out}} = 0.041$. We model the lobes as triaxial ellipsoids of dimensions approximately $22 \times 20 \times 7$ and $14 \times 14 \times 10$ km³ (ref. ¹), leading to a total radius $R_{\text{tot}} = 18$ km and inner mass $m_{\text{in}} = (1.61 + 1.03) \times 10^{18}$ g = 2.64×10^{18} g for a density of $\rho = 1$ g cm^{−3} (see Methods for other densities). Secular collisions occur only for sufficiently large critical inclination and beyond a certain initial separation α_{coll} , which overcomes LK quenching (see equation (12) and Methods). Non-secular collisions will dominate over secular collisions beyond a transitional separation α_{t} :

$$\alpha_{\text{t}} = 3^{1/3} \left[\frac{128}{135} \frac{(1 - e_{\text{out}}^2)}{\left(1 + \frac{2\sqrt{2}}{3} e_{\text{out}}\right)^2} \left(\frac{M_{\odot}}{m_{\text{in}}}\right)^{1/3} \frac{R_{\text{tot}}}{a_{\text{out}}} \right]^{1/4} \quad (1)$$

¹Physics Department, Technion–Israel Institute of Technology, Haifa, Israel. ²Institut für Astronomie und Astrophysik, Eberhard Karls Universität Tübingen, Tübingen, Germany.

✉e-mail: eugeneg@campus.technion.ac.il

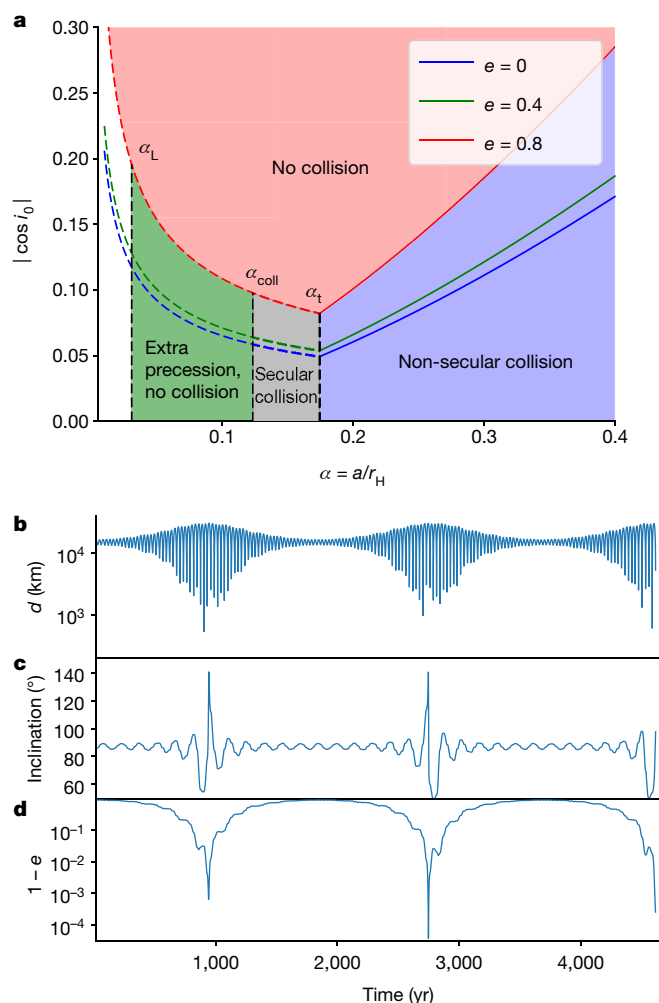


Fig. 1 | Roadmap to collisions of MU₆₉. **a**, The orbital evolution of MU₆₉ in initial separation–initial inclination space. The initial eccentricity, e , is 0 (blue), 0.4 (green) and 0.8 (red). Solid lines show the condition for non-secular collisions (equation (14)) with unbound eccentricity. Dashed lines show the conditions for a secular collision (equation (4)) with deterministic eccentricity. The different domains are as follows: white, LK oscillations are completely quenched and the eccentricity is constant, $\alpha < \alpha_L$ (equation (9)); green, the eccentricity is excited but below e_{coll} , and collisions are avoided, $\alpha_L < \alpha < \alpha_{\text{coll}}$ (equation (12)); grey, secular evolution can lead to a collision, $\alpha_{\text{coll}} < \alpha < \alpha_t$ (equation (1)); blue, non-secular perturbations dominate and lead to a collision; red, the initial inclination is too low to induce a collision. **b–d**, Time evolution of the instantaneous distance, inclination and eccentricity of an individual orbit with initial Keplerian elements: semi-major axis $a = 0.3r_H$, eccentricity $e = 0.1$, inclination $i = 86^\circ$, argument of periape $\omega = 0$, argument of ascending node $\Omega = \pi/4$ and mean anomaly $M = 0$. The outer binary is set at $\omega_{\text{out}} = \Omega_{\text{out}} = 0$ and $M_{\text{out}} = -\pi/4$.

In our case, $\alpha_t \approx 0.174$. Figure 1b demonstrates the separation in the non-secular regime before the collision. During the high-eccentricity phase, there are about 10 cycles where the instantaneous separation drops below 10^3 km. A collision occurs during the third LK cycle after about 4,600 yr. The mutual inclination flips its orientation during the high-eccentricity peak of the LK cycle (Fig. 1c). The eccentricity is essentially unbound and a collision eventually occurs (Fig. 1d).

To explore in detail the overall evolution and statistics of KBBs in the chaotic non-secular regime, we defer to detailed N -body simulations, which provide us with the probability for collisions and the post-collision characteristics. We use the publicly available code REBOUND¹⁸ with the IAS15¹⁹ integrator (see Methods for details and stopping conditions). We integrate four sets of initial conditions in

the non-secular regime. The first three sets have initial separations of $\alpha = 0.2, 0.3$ and 0.4 , and the fourth set has uniformly sampled separations in $\alpha = [0.2, 0.4]$. The orbital angles are sampled uniformly. The mutual inclination of observed binaries is cosine-uniform⁹, and thus we follow cosine-uniform sampling with a cut-off at $|\cos i| \leq 0.4$ (lower inclinations cannot lead to a collision). For each case, we run 250 simulations (except for $\alpha = 0.2$, for which we run 200 simulations and use $|\cos i| \leq 0.3$), each up to 5×10^4 yr.

Figure 2 shows the cumulative distribution function of various parameters of the colliding orbits. Both the closest-approach distance $q = a(1-e)/R_{\text{tot}}$ (Fig. 2a) and the final inclinations (Fig. 2c) at collision are consistent with a uniform distribution (in $\cos i$) between 40° and 140° , suggesting that the orbits are indeed chaotic and in the non-secular regime, as expected. Most orbits induce collisions after about a few thousand years (Fig. 2b). The mean collision time increases with increasing separation. The velocity at impact is comparable to the escape velocity with a very small dispersion, consistent with a gentle collision²⁰ (Fig. 2d).

We find the overall merger fractions of wide binaries to be around 12%–18% (see Extended Data Table 1), which is roughly consistent with the observed 10%–25% occurrence of contact binaries for the cold classical belt²¹. Most mergers occur for initially high inclinations, as expected. About 1%–3% of all wide binaries produce highly oblique contact binaries ($i = 80^\circ$ – 100°), consistent with the observed high obliquity of MU₆₉, and providing predictions that can be verified by future KBO observations. There is little dependence on the underlying distribution of α , and merger rates are bounded between minimal and maximal values of 12% (for $\alpha = 0.2$) and 18% (for $\alpha = 0.4$). Moreover, in a collisional environment²² the binary orbits can be perturbed such that originally low-inclination orbits become highly inclined and become subject to semi-secular evolution, forming contact binaries; the quoted formation rates are thus lower limits to the total fraction of contact binaries formed through this process.

The non-merging systems continue to evolve quasi-periodically. On longer timescales, three-body encounters are expected to shape the populations of KBBs^{3,23}. Exchange interactions can drive the binaries into equal masses²⁴, and the loose nature of the binaries can result in evaporation (Heggie’s law)²⁵. There are only a handful of KBBs beyond $a \geq 0.05r_H$ with either prograde or retrograde orbits that are not highly inclined (see figure 1 of ref. ²⁶), whereas the widest known binary, 2001 QW₃₂₂, with $a \approx 0.2r_H$, is expected to disrupt within a billion years¹⁶.

To test the feasibility of the semi-secular collision origin of MU₆₉, we also need to account for the observed spin period of MU₆₉. Angular momentum conservation enables us to find the resulting spin period depending on the impact angle and the primordial spins of each component. The final impact parameter at collision (which corresponds to an impact angle; see Methods) is uniformly distributed, and thus our model can robustly produce a wide range of possible final rotation periods, without any fine-tuned modelling of the composition and density of MU₆₉, thus also alleviating the angular momentum problem of other models¹.

Figure 3a shows the outcome of a collision at a 40° impact angle with high-material-strength composition, which reproduces the shape of MU₆₉. Low- or medium-strength materials result in a deformed shape and are thus ruled out. If the density of MU₆₉ is halved compared to the fiducial 1 g cm^{-3} value (as suggested by ref. ¹), the escape velocity v_{esc} —at which typical collisions occur—is lower, and thus using medium-strength-material parameters also produces an undeformed shape. Random collisions—even at relative velocities as low as $10v_{\text{esc}}$ —destroy or heavily deform binaries with high-strength-material composition; they are likewise ruled out (see Methods). Figure 3b shows the expected spin-period dependence on the impact angle. An impact angle of about 40° reproduces the observed spin period (see Methods) for initially non-spinning objects. Taking a typical initial spin period of about 10 h with random orientations extends the range of plausible

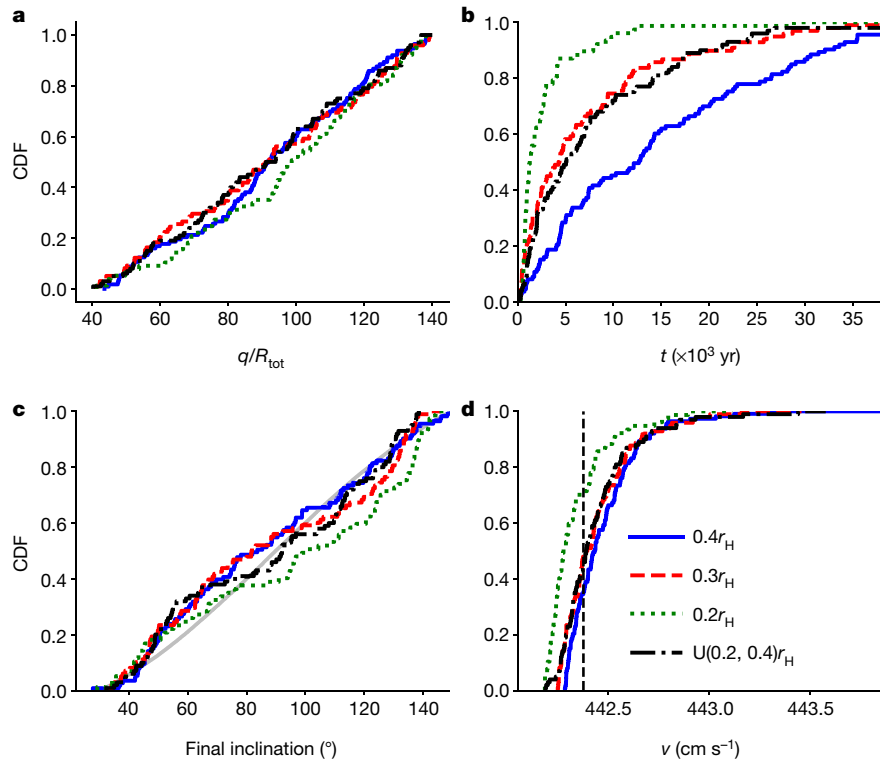


Fig. 2 | Cumulative distributions of the impact characteristics. The cumulative distribution functions (CDFs) are for $\alpha = 0.4$ (solid blue), 0.3 (dashed red), 0.2 (dotted green) and uniform in $(0.2, 0.4)$ (dash-dotted black). **a**,

Pericentre q/R_{tot} . **b**, Time of collision. **c**, Final inclination at impact. The shaded grey line is a uniform cumulative distribution in $\cos i$ in the range 30° – 150° . **d**, Velocity at impact. The vertical black dashed line is the escape velocity, v_{esc} .

impact angles to about 20° and 70° , for the maximally aligned and anti-aligned configurations, respectively. Smoothed particle hydrodynamics (SPH) collision simulations agree with our simplified estimate and support our assumptions of undeformed, rigid bodies when modelled with high-strength material parameters, or else medium-high strength parameters if the density and impact velocity are slightly lower (see ref.²⁷ and Methods for details).

Together, our dynamical and post-collisional modelling yields a coherent picture for the origin of MU₆₉ from an ultrawide KBO binary. Such wide KBB progenitors could be a natural byproduct of KBO and KBB evolution in the early Solar System^{3,28,29}. It is most probable that the characteristics of MU₆₉ are not unique, and that secular or semi-secular evolution plays a major role in the evolution of many KBBs and in the

production of low-velocity collisions between individual KBB components. In fact, modelling of the Pluto–Charon system also suggests a low-velocity impact origin³⁰. Moreover, given the high obliquity of the Pluto–Charon system, it is possible that it also originated from an initially wide binary and followed a secular or semi-secular evolution, similar to MU₆₉. Similar evolutionary scenarios might also apply to the evolution of other contact binaries such as (139775) 2001 QG298 (ref.²), as well as moons and exo-moons, as all of these form hierarchical triple systems with their host star.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2194-z>.

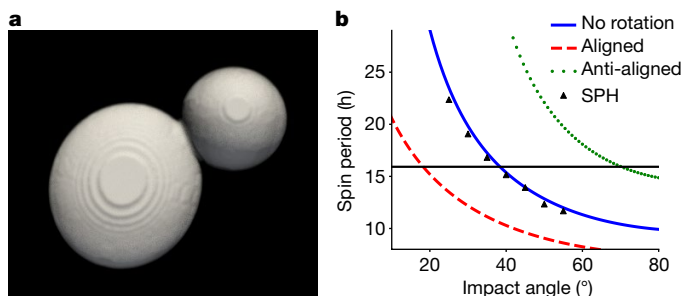


Fig. 3 | Shape and spin period of MU₆₉. **a**, Final collision outcome at an impact angle of 40° with high-material-strength composition (see Methods). **b**, Spin period as a function of the impact angle for MU₆₉. The horizontal line is the observed period, 15.92 h. The solid blue line indicates initially non-rotating progenitors. The dashed red line indicates two initially aligned rotating progenitors with a period of 10 h, and the dotted green line indicates an anti-aligned configuration. The black crosses are the results obtained from SPH simulations (see Methods).

1. Stern, S. A. et al. Initial results from the New Horizons exploration of 2014 MU₆₉, a small Kuiper belt object. *Science* **364**, eaaw9771 (2019).
2. Lacerda, P. A change in the light curve of Kuiper belt contact binary (139775) 2001 QG₂₉₈. *Astron. J.* **142**, 90–98 (2011).
3. Goldreich, P., Lithwick, Y. & Sari, R. Formation of Kuiper belt binaries by dynamical friction and three-body encounters. *Nature* **420**, 643–646 (2002).
4. Richardson, D. C. & Walsh, K. J. Binary minor planets. *Annu. Rev. Earth Planet. Sci.* **34**, 47–81 (2006).
5. Perets, H. B. & Naoz, S. Kozai cycles, tidal friction, and the dynamical evolution of binary minor planets. *Astrophys. J. Lett.* **699**, 17–21 (2009).
6. Antonini, F. & Perets, H. B. Secular evolution of compact binaries near massive black holes: gravitational wave sources and other exotica. *Astrophys. J.* **757**, 27–40 (2012).
7. Grishin, E., Perets, H. B. & Fragione, G. Quasi-secular evolution of mildly hierarchical triple systems: analytics and applications for GW sources and hot Jupiters. *Mon. Not. R. Astron. Soc.* **481**, 4907–4923 (2018).
8. Grishin, E., Perets, H. B., Zenati, Y. & Michaely, E. Generalized Hill-stability criteria for hierarchical body systems at arbitrary inclinations. *Mon. Not. R. Astron. Soc.* **466**, 276–285 (2017).

9. Naoz, S., Perets, H. B. & Ragozzine, D. The observed orbital properties of binary minor planets. *Astrophys. J.* **719**, 1775–1783 (2010).
10. Lidov, M. L. The evolution of orbits of artificial satellites of planets under the action of gravitational perturbations of external bodies. *Planet. Space Sci.* **9**, 719–759 (1962).
11. Kozai, Y. Secular perturbations of asteroids with high inclination and eccentricity. *Astron. J.* **67**, 591–598 (1962).
12. Naoz, S. The eccentric Kozai–Lidov effect and its applications. *Annu. Rev. Astron. Astrophys.* **54**, 441–489 (2016).
13. Fabrycky, D. & Tremaine, S. Shrinking binary and planetary orbits by Kozai cycles with tidal friction. *Astrophys. J.* **669**, 1298–1315 (2007).
14. Porter, S. B. & Grundy, W. M. KCTF evolution of trans-Neptunian binaries: connecting formation to observation. *Icarus* **220**, 947–957 (2012).
15. Veillet, C. et al. The binary Kuiper-belt object 1998 WW31. *Nature* **416**, 711–713 (2002).
16. Petit, J. M. et al. The extreme Kuiper belt binary 2001 QW₃₂₂. *Science* **322**, 432–434 (2008).
17. Luo, L., Katz, B. & Dong, S. Double-averaging can fail to characterize the long-term evolution of Lidov–Kozai cycles and derivation of an analytical correction. *Mon. Not. R. Astron. Soc.* **458**, 3060–3074 (2016).
18. Rein, H. & Liu, S.-F. REBOUND: an open-source multi-purpose *N*-body code for collisional dynamics. *Astron. Astrophys.* **537**, A128 (2012).
19. Rein, H. & Spiegel, D. S. IAS15: a fast, adaptive, high-order integrator for gravitational dynamics, accurate to machine precision over a billion orbits. *Mon. Not. R. Astron. Soc.* **446**, 1424–1437 (2015).
20. McKinnon, W. B. et al. The solar nebula origin of (486958) Arrokoth, a primordial contact binary in the Kuiper belt. *Science* **367**, eaay6620 (2020).
21. Thirouin, A. & Sheppard, S. S. Light curves and rotational properties of the pristine cold classical Kuiper belt objects. *Astron. J.* **157**, 228–247 (2019).
22. Parker, A. H. & Kavelaars, J. J. Collisional evolution of ultra-wide trans-Neptunian binaries. *Astrophys. J.* **744**, 139–152 (2012).
23. Perets, H. B. Binary planetesimals and their role in planet formation. *Astrophys. J. Lett.* **727**, 3 (2011).
24. Funato, Y., Makino, J., Hut, P., Kokubo, E. & Kinoshita, D. The formation of Kuiper belt binaries through exchange reactions. *Nature* **427**, 518–520 (2004).
25. Heggie, D. C. Binary evolution in stellar dynamics. *Mon. Not. R. Astron. Soc.* **173**, 729–787 (1975).
26. Grundy, W. et al. Mutual orbit orientations of transneptunian binaries. *Icarus* **334**, 62–78 (2019).
27. Schäfer, C. et al. A smooth particle hydrodynamics code to model collisions between solid, self-gravitating objects. *Astron. Astrophys.* **590**, A19 (2016).
28. Goldreich, P., Lithwick, Y. & Sari, R. Planet formation by coagulation: a focus on Uranus and Neptune. *Annu. Rev. Astron. Astrophys.* **42**, 549–601 (2004).
29. Nesvorný, D., Li, R., Youdin, A. N., Simon, J. B. & Grundy, W. M. Trans-Neptunian binaries as evidence for planetesimal formation by the streaming instability. *Nat. Astron.* **3**, 808–812 (2019).
30. Canup, R. M. A giant impact origin of Pluto–Charon. *Science* **307**, 546–550 (2005).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Lidov–Kozai secular evolution

Let us consider the evolution of a binary KBO due to LK secular evolution. Let the inner binary start with an initial separation of $a_0 = \alpha r_H$, eccentricity e_0 and mutual inclination i_0 . The Hill radius is $r_H = a_{\text{out}}(m_{\text{in}}/3M_{\odot})^{1/3}$. The most stable orbit is around $\alpha_H \approx 0.4$ (ref. ⁸). The minimal eccentricity required for collision is

$$e_{\text{coll}} = 1 - \frac{R_{\text{tot}}}{\alpha r_H} \quad (2)$$

Using the standard LK formula, the maximal eccentricity is

$$e_{\text{max}}^2 = 1 - \frac{5}{3}(1 - e_0^2)\cos^2 i_0 \quad (3)$$

For a collision occur, we require that $e_{\text{max}} \geq e_{\text{coll}}$, which yields the critical inclination i_0^c

$$\cos i_0^c \approx \sqrt{\frac{6}{5(1 - e_0^2)} \frac{R_{\text{tot}}}{\alpha r_H}} = \frac{0.037}{\sqrt{1 - e_0^2}} \left(\frac{\alpha}{0.3} \right)^{-1/2} \quad (4)$$

such that collisions occur for $|\cos i_0| \leq |\cos i_0^c|$. Here, $R_{\text{tot}}/r_H \ll 1$ has been expanded to linear order. The probability for collision can be expressed in terms of integrating over the distribution function $f_{a,\theta}(a, \theta)$ and where $\theta \equiv \cos i$. For uniform independent distribution in $\cos i$, as inferred from KBO observations⁹, the probability is

$$P(e_0) = \int_{a_{\text{min}}}^{a_{\text{max}}} f_a(a) \theta_c(a) da \quad (5)$$

where $\theta_c(a)$ is given by equation (4).

Inclination angle at impact. From the conservation of $\sqrt{1 - e^2} \cos i = \text{const} \equiv j_z$ the inclination at impact is

$$\cos i_{\text{coll}} = \frac{\sqrt{1 - e_0^2} \cos i_0}{\sqrt{1 - e_{\text{coll}}^2}} = \sqrt{1 - e_0^2} \cos i_0 \sqrt{\frac{\alpha r_H}{2R_{\text{tot}}}} \quad (6)$$

To find the impact angle i_{coll} , we invert equation (6):

$$i_0 = \arccos \left(\cos i_{\text{coll}} \sqrt{\frac{2R_{\text{tot}}}{\alpha r_H}} \right) \quad (7)$$

To find the observed obliquity $i_{\text{coll}} = 98^\circ$, the critical inclination is at least $|\cos i_0| \leq 0.00585$, ($89.33^\circ \leq i_0 \leq 90.66^\circ$) for $\alpha = 0.1$, which is unlikely. Moreover, for $\alpha \leq 0.1$, collisions are unlikely regardless of the initial inclination, owing to the effects of oblateness, as shown below.

Effects of oblateness. Small KBOs might not have spherical shapes, in which case their gravitational potential is not spherical. Such a configuration induces extra precession on the orbit which can considerably affect the secular evolution. The leading term is encapsulated in a dimensionless parameter J_2 , which is related to ratio of the axes, or the polar and equatorial radii of the bodies³¹. Planets are mostly spherical, and their deviation is small— $J_2 \approx 10^{-3}$ for Earth and around $J_2 = 0.014$ for Jupiter—and is related to the flattening of the planets induced by their rotations. In the case of the components of MU₆₉, the objects are highly non-spherical and J_2 could be large. Using the principal moments of inertia of an oblate spheroid, we have $J_2 = [1 - (c/a)^2]/5$, which is around $J_2 \approx 0.18$ for the primary component and $J_2 \approx 0.1$ for the secondary component.

The additional precession may quench the LK oscillations if it is too strong. To quantify the effects of the additional precession we can

define a dimensionless quantity, ε_{rot} , that measures the ratio between the LK-induced and the oblateness-induced precessions^{32,33}

$$\varepsilon_{\text{rot}} = \frac{3}{2} J_2 \frac{m_{\text{in}}}{m_{\text{out}}} \frac{a_{\text{out}}^3 (1 - e_{\text{out}}^2)^{3/2} R_1^2}{(\alpha r_H)^5} \quad (8)$$

Setting $\varepsilon_{\text{rot}} = 3/2$ leads to the definition of the Laplace radius^{33,34} in terms of the Hill radius, $r_L \equiv \alpha_L r_H$ where

$$\alpha_L = \left(J_2 \frac{m_{\text{in}}}{m_{\text{out}}} \right)^{1/5} \frac{(a_{\text{out}}^3 R_1^2)^{1/5}}{r_H} (1 - e_{\text{out}}^2)^{3/10} = 0.03 \left(\frac{J_2}{0.2} \right)^{1/5} \left(\frac{R_1}{11 \text{ km}} \right)^{2/5} \quad (9)$$

leading to $\varepsilon_{\text{rot}} = 1.5(\alpha/\alpha_L)^{-5}$. It has been previously shown that the maximal eccentricity attained is given by the implicit expression for $\cos i_0 = 0$ (their equation 50):

$$\frac{\varepsilon_{\text{rot}}}{3} \left[\frac{1}{(1 - e_{\text{max}}^2)^{3/2}} - 1 \right] = \frac{9}{8} e_{\text{max}}^2 \quad (10)$$

Expanding in $\varepsilon_{\text{rot}} \ll 1$ and $e_{\text{max}}^2 \approx 1$

$$e_{\text{max}} \approx 1 - \frac{2}{9} \varepsilon_{\text{rot}}^{2/3} \quad (11)$$

A collision can occur only if $e_{\text{coll}} < e_{\text{max}}$ or if

$$\alpha_{\text{coll}} > \left(\frac{3}{2} \right)^{2/7} \alpha_L^{10/7} \left(\frac{2}{9} \frac{r_H}{R_{\text{tot}}} \right)^{3/7} \approx 0.12 \quad (12)$$

Note that a similar analysis can be performed for tidal distortions or relativistic corrections³². In this case, they are much weaker than the rotational effects.

Non-secular Lidov–Kozai evolution

In the previous section we considered the evolution due to secular LK evolution. In the semi-secular (semi-LK) regime^{6,7}, short-term fluctuations can substantially change the evolution. In the following we discuss the overall effects of such short-term perturbations. The strength of the perturbations is encapsulated in the single averaging parameter^{7,17}:

$$\varepsilon_{\text{SA}} \equiv \left(\frac{a_1}{b_{\text{out}}} \right)^{3/2} \sqrt{\frac{m_{\text{out}}}{m_{\text{in}}}} = \frac{\alpha^{3/2}}{\sqrt{3}(1 - e_{\text{out}}^2)^{3/2}} \approx 0.1 \left(\frac{\alpha}{0.3} \right)^{3/2} \quad (13)$$

One important quantity is the (averaged) z angular momentum $\bar{j}_z = \sqrt{1 - e_0^2} \cos i_0$, assuming that i_0 and e_0 have their mean value.

In this case, \bar{j}_z is no longer conserved, but its value averaged over the outer orbit, \bar{j}_z , is conserved. The eccentricity of the orbit becomes unbound once the fluctuation in \bar{j}_z ($\Delta \bar{j}_z$) is larger than its initial value, namely $\Delta \bar{j}_z > \bar{j}_z$. The fluctuation has been estimated analytically^{7,17}, and can be used to show that the eccentricity is unbound if

$$\cos i_0 \sqrt{1 - e_0^2} \lesssim \frac{9}{8} \varepsilon_{\text{SA}} \approx 0.118 \left(\frac{\alpha}{0.3} \right)^{3/2} \quad (14)$$

where $\tilde{\varepsilon}_{\text{SA}} = \varepsilon_{\text{SA}}(1 + 2\sqrt{2}e_{\text{out}}/3) \approx 1.039\varepsilon_{\text{SA}}$ has been defined for convenience.

The width of the non-secular semi-LK regime increases with α , whereas the width of the secular LK regime decreases with increasing α . Comparing equation (4) and equation (14) yields the transitional separation α_t found in equation (1).

Spin period

There is little evidence for structural changes of MU₆₉ since its formation, and the spin period is believed to be primordial¹. In our model, the collision is gentle and occurs at relatively low velocities

Article

($v_{\text{esc}} = 442.4 \text{ cm s}^{-1}$ for our nominal density and $v_{\text{esc}} = 3.128 \text{ cm s}^{-1}$ for the lower density of 0.5 g cm^{-3} assumed in ref. ¹), such that almost any impact parameter (or impact angle) is allowed. Therefore, to obtain the observed spin period, we can use the standard arguments of angular momentum conservation and derive the impact parameter (or impact angle) that yields the desired spin rate.

Consider triaxial ellipsoidal bodies i with masses m_i and axes $a_i \geq b_i \geq c_i$, with $i = 1, 2$. We assume that the major axes a_i are parallel, similar to the observed object, and that the collision occurs in parallel with the major axes. The largest moment of inertia is $I_3^{(i)} = m_i(a_i^2 + b_i^2)/5$.

After the collision the distance between the centre of masses of the joint body and each centre of the ellipsoid is r_i . Then the principal moment of inertia of the joint body is

$$I_3^{\text{tot}} = I_3^{(1)} + I_3^{(2)} + m_1 r_1^2 + m_2 r_2^2 = \sum_{i=1}^2 \frac{m_i}{5} (a_i^2 + b_i^2 + 5r_i^2)$$

Now, the ellipsoids collide with relative velocity v_{esc} and impact parameter b . The orbital angular momentum is $L_z = \mu b v_{\text{esc}}$, where $\mu = m_1 m_2 / (m_1 + m_2)$ is the reduced mass. If the two bodies are non-rotating, then the joint angular frequency is

$$\Omega = \frac{L_z}{I_3^{\text{tot}}} = \frac{5\mu b v_{\text{esc}}}{m_1(a_1^2 + b_1^2 + 5r_1^2) + m_2(a_2^2 + b_2^2 + 5r_2^2)} \quad (15)$$

If the individual bodies are rotating around the z axis with frequencies Ω_i , the additional angular momentum of each body is $I_3^{(i)} \Omega_i$ for $i = 1, 2$, and thus equation (15) becomes

$$\Omega = \frac{5\mu b v_{\text{esc}} + m_1(a_1^2 + b_1^2)\Omega_1 + m_2(a_2^2 + b_2^2)\Omega_2}{m_1(a_1^2 + b_1^2 + 5r_1^2) + m_2(a_2^2 + b_2^2 + 5r_2^2)} \quad (16)$$

For an impact angle θ , the distance of the point of contact to the centre of each ellipsoid is $\xi_i = \sqrt{a_i^2 \cos^2 \theta + b_i^2 \sin^2 \theta}$. The impact parameter is related to the impact angle by $\sin \theta = b / (\xi_1 + \xi_2) = b/d$. The distances from the centre of mass are $r_1 = m_2 d / (m_1 + m_2)$ and $r_2 = m_1 d / (m_1 + m_2)$ and the spin rate is

$$\Omega = \frac{5\mu d v_{\text{esc}} \sin \theta + m_1(a_1^2 + b_1^2)\Omega_1 + m_2(a_2^2 + b_2^2)\Omega_2}{5\mu d^2 + m_1(a_1^2 + b_1^2) + m_2(a_2^2 + b_2^2)} \quad (17)$$

Figure 3 shows the spin-period dependence on the impact angle for the typical parameters of MU₆₉. The spin period is $P = 2\pi/\Omega$, where Ω is given by equation (17), and when there is no internal rotation, $\Omega_1 = \Omega_2 = 0$. We see that an impact angle of about 40° gives the observed spin period. We have performed hydrodynamical simulations based on the code of ref. ²⁷ that qualitatively agree with our assumptions and produce similar results. Typical classical KBOs could have primordial spin periods³⁵ with comparable contributions to the angular momentum budget. Recently, it was found²¹ that the mean cold classical KBO spin period is $9.48 \pm 1.53 \text{ h}$. Generally, there is no reason for the spin vectors of each body to be correlated, so on average the contribution is zero. In extreme cases, the spin vector of both objects could be aligned or anti-aligned with the orbital angular momentum. In these cases, a large range of impact angles and spin configurations are possible, resulting in the observed spin period after the collision. For a typical period of 10 h, the impact angle is about 20° in the aligned case, and about 70° in the anti-aligned case.

N-body stopping conditions and tests

We impose a stopping condition that the distance between the two bodies is less than their mutual radius. During the non-secular highly eccentric passage, the change in the pericentre q is much faster than the

inner orbital period (this is the definition of the non-secular regime), and hence the orbital elements are not reliable at this stage. Once the simulation stops it records the orbital elements at impact, which we use for our statistics, but these are not involved in the stopping condition. From the output we know the closest approach at impact. We have tested the stopping condition by varying it to be slightly smaller or larger than the q found in the first run. Indeed, when the stopping condition was below q the objects did not collide and the code continued running. We thereby concluded that the collision is physical and reliable.

Extended Data Table 1 shows the merger fractions from the simulations. The total merger fraction f_i is the total number of mergers divided by the initial number of runs, multiplied by the relative fractions of the inclination distribution, assuming that no mergers occur outside the sampled inclination distribution. The fraction f_{80-100} is calculated in the same way, except that the only mergers considered are those where the mutual inclination during the merger is within the designated boundaries of 80° – 100° . For example, for $\alpha = 0.2$, the merger fraction is $78/200 = 0.39$. Multiplied by the range of the inclination distribution, $f_i = 0.39 \times 0.3 \approx 0.12$ and the high-obliquity merger fraction is $f_{80-100} = 9/200 \times 0.3 \approx 0.014$. For $\alpha = 0.3$, the merger fraction is $99/250 = 0.396$. Multiplied by the range of the inclination distribution, $f_i = 0.396 \times 0.4 = 0.158$ and the high-obliquity merger fraction is $f_{80-100} = 12/250 \times 0.4 \approx 0.019$.

Impact modelling

We perform hydrodynamical collision simulations using our SPH code²⁷, which treats self-gravity, gas, fluid, elastic and plastic solid bodies that have a material strength, including a porosity and fracture model that can be applied for small-body collisions^{36,37}. In order to treat numerical rotational instabilities, a tensorial correction scheme³⁸ is implemented. The miluphCUDA code is implemented with CUDA, and runs on graphics processing units (GPUs), with an improvement of approximately one to two orders of magnitude for a single GPU compared to a single central processing unit (CPU). The code has previously been successfully applied to several studies involving impact processes^{36,37,39–47}.

For the porosity treatment, we implement the P – α model^{48,49}, in which the pores are much smaller than the spatial resolution and cannot be modelled explicitly. Here, the total change in the volume depends both on the compaction or collapse of the pore space and on the compression of the solid material that constitutes the matrix. The dependence is expressed in terms of a porous material pressure P and density ρ as $P/\rho = P_s/\rho_s$, where P_s and ρ_s are the pressure and density of the solid matrix material, respectively. The distention parameter $\bar{\alpha} = \rho_s/\rho$ is the ratio between the solid matrix material and the porous material densities, and relates to the porosity ψ via $\psi = 1 - 1/\bar{\alpha}$. For the solid matrix material we use the Tillotson equation of state (EOS) parameters⁵⁰ with a reduced bulk modulus of $A = 2.67 \times 10^8 \text{ Pa}$ (the leading term in the EOS) to take into account the smaller elastic wave speeds in porous materials compared to solid materials, consistent with previous work⁵¹. Our matrix density is chosen to be 2 g cm^{-3} , about the same as that used previously⁵², which leads to a 50% porosity for our fiducial bulk density for MU₆₉, 1 g cm^{-3} . The matrix density and the initial porosity are both in rough agreement with what might be expected from an object of this origin and size range. In particular, the former constrains the rock–ice mass ratio to be about 3–4 (depending on the exact choice of silicate grain density), which could be compatible with this type of KBO^{53–56}. However, we note that given the uncertainties involved, we seek only to obtain a rough estimate of the density that will permit us to test our working hypothesis. We then also run simulations with 75% porosity and half the previous bulk density to establish the qualitative differences between these two setups.

For collisions between small porous bodies, compressibility is limited by the crush curve for $\bar{\alpha}$ for typical pressures, instead of by the Tillotson

EOS parameters. We thereby choose three sets of crush-curve parameters⁵², using a simple quadratic crush curve⁵⁷:

$$\bar{\alpha} = 1 + (\bar{\alpha}_0 - 1) \frac{(P_s - P)^2}{(P_s - P_e)^2} \quad (18)$$

where $\bar{\alpha}_0 = 2$, P_e is the transition pressure between the elastic and plastic regimes and P_s is the pressure of full compaction. Both P_e and P_s are listed in Extended Data Table 2. As ref. ⁵² treats comet 67P/Churyumov–Gerasimenko, which belongs to a class of much smaller and active objects, we assume MU₆₉ is probably fluffier and more porous. Hence, our low-strength crush-curve values correspond to the previous high-strength values⁵², and taking the same modelling approach we then increment the parameters in each subsequent model by one order of magnitude.

Fracture and brittle failure are treated using the Grady and Kipp fragmentation prescription^{58–60}, which is based on randomly distributed flaws in the material following a Weibull distribution with material-dependent parameters. The lowest activation threshold strain, κ , derived from the Weibull distribution, is given by $\kappa = kV^{-1/m}$, where V is the volume of the brittle material and k and m are the material-dependent Weibull parameters. We adopt $m = 9.5$ for pressure-dependent failure⁶¹. The volume is calculated given the dimensions of the MU₆₉ binary. From the material strength parameters K and G , Young's modulus E may be calculated as $E = (9KG)/(3K + G)$. Here $K = 2.67 \times 10^8$ Pa, which is the leading term in the Tillotson EOS, and $G = 1.6 \times 10^8$ Pa. Finally, for undamaged material, $\kappa = Y_T/E$, where Y_T is the tensile strength given in table 30 of ref. ⁵². k may thus be extracted and is $k = 10^{47}$, 2×10^{39} and 2×10^{28} m⁻³ for the low-, medium- and high-strength-material setups, respectively. Damage accumulates when the local tensile strain reaches the activation threshold of a flaw.

For the plasticity model we use a pressure-dependent yield strength⁶² following the implementation of ref. ⁶¹. The yield stress Y_i is different for damaged and intact material. For intact material, the yield stress is $Y_i = Y_0 + \mu_i P / (1 + \mu_i P / (Y_M - Y_0))$, where Y_0 is the cohesion (again see table 30 of ref. ⁵²), μ_i is the coefficient of friction and Y_M is the shear strength at $P = \infty$. We adopt $\mu_i = 1.5$ (ref. ⁶¹) and a typical $Y_M = 1.5 \times 10^9$ Pa (ref. ⁶⁰), which is appropriate for an object composed of ice, rock and organics. For $P = 0$, we recover the pressure-independent form $Y_i = Y_0$. For damaged material the yield stress is $Y_d = \mu_d P$, where μ_d is the coefficient of friction of the damaged material. Here we take $\mu_d = 0.6$, following ref. ⁶¹, and thus fully damaged particles still undergo some shear stress.

Extended Data Fig. 1 shows additional results of our simulated impacts. We obtain the rotation period of MU₆₉ using the nominal density of 1 g cm⁻³ only when using the high-strength-material parameters. Medium-strength (Extended Data Fig. 1a) or low-strength (Extended Data Fig. 1b) materials deform MU₆₉ and do not produce the observed shape of a gently merged contact binary. If the nominal density is halved (0.5 g cm⁻³), the impact velocity v_{esc} is lower, which produces less deformation in our simulations. Even medium-strength material parameters generate a gently merged contact binary for virtually all impact angles (Extended Data Fig. 1c, d). Here, we used the 55° impact angle, for which the observed spin period of MU₆₉ is approximately obtained. In Extended Data Fig. 2 the shape is considerably deformed after the collision if the impact velocity is larger than $v = 10v_{\text{esc}}$, using high-strength material parameters. The same velocity with weaker material parameters leads to a complete disruption of MU₆₉.

Our simulations were performed for a grid of impact angles, assuming pre-alignment of the two lobes. Simulating higher impact angles and low- and medium-strength-material compositions causes the two lobes to interact in other ways: in some cases they hit and create contact craters and then roll on top of each other; in other cases they collide, bouncing off each other instead of rolling, and then return to re-collide following a (now) shorter orbital period. These formation channels may also generate compatible shapes but, so far, not the exact rotation

period of MU₆₉. A full investigation of the collision phase space must also include the initial self-rotation of each lobe in addition to inclined hits. This requires a huge collision phase space, exceeding the scope of this work, and necessitates a dedicated hydrodynamical study. Preliminary results (in preparation) indicate that such an approach may yield more channels through which the unique orientation of MU₆₉ might be generated, besides the successful cases for pre-aligned binary components that are shown here.

Our standard resolution is 5×10^5 SPH particles. We have additionally preformed simulations with 10^5 and 2.5×10^5 particles. Test simulations were performed on the TAMNUN GPU cluster at the Technion Institute in Israel and production runs on the bwForCluster BinAC at the University of Tübingen, Germany.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

- Murray, C. D. & Dermott, S. F. *Solar System Dynamics* (Cambridge Univ. Press, 1999).
- Liu, B., Muñoz, D. J. & Lai, D. Suppression of extreme orbital evolution in triple systems with short-range forces. *Mon. Not. R. Astron. Soc.* **447**, 747–764 (2015).
- Grishin, E., Lai, D. & Perets, H. B. Chaotic quadruple secular evolution and the production of misaligned exomoons and warm Jupiters in stellar multiples. *Mon. Not. R. Astron. Soc.* **474**, 3547–3556 (2018).
- Tremaine, S., Touma, J. & Namouni, F. Satellite dynamics on the Laplace surface. *Astron. J.* **137**, 3706–3717 (2009).
- Thirouin, A., Noll, K. S., Ortiz, J. L. & Morales, N. Rotational properties of the binary and non-binary populations in the trans-Neptunian belt. *Astron. Astrophys.* **569**, A3 (2014).
- Wandel, O. J., Schäfer, C. M. & Maindl, T. I. Collisional fragmentation of porous objects in planetary systems. In *Proc. 1st Greek–Austrian Workshop Extrasolar Planetary Systems* (eds Maindl, T. I., Varvoglis, H. & Dvorak, R.) 225–242 (2017).
- Haghighipour, N., Maindl, T. I., Schäfer, C. M. & Wandel, O. J. Triggering the activation of main-belt comets: the effect of porosity. *Astrophys. J.* **855**, 60 (2018).
- Speith, R. *Improvements of the Numerical Method of Smoothed Particle Hydrodynamics*. Habilitation thesis, Univ. of Tübingen (2006).
- Dvorak, R., Maindl, T. I., Burger, C., Schäfer, C. & Speith, R. Planetary systems and the formation of habitable planets. *Nonlinear Phenom. Complex Syst.* **18**, 310–325 (2015).
- Maindl, T. I. et al. Impact induced surface heating by planetesimals on early Mars. *Astron. Astrophys.* **574**, A22 (2015).
- Haghighipour, N., Maindl, T. I., Schäfer, C., Speith, R. & Dvorak, R. Triggering sublimation-driven activity of main belt comets. *Astrophys. J.* **830**, 22 (2016).
- Schäfer, C. M. et al. Numerical simulations of regolith sampling processes. *Planet. Space Sci.* **141**, 35–44 (2017).
- Burger, C., Maindl, T. I. & Schäfer, C. M. Transfer, loss and physical processing of water in hit-and-run collisions of planetary embryos. *Celestial Mech. Dyn. Astron.* **130**, 2 (2018).
- Malamud, U., Perets, H. B., Schäfer, C. & Burger, C. Moonfalls: collisions between the Earth and its past moons. *Mon. Not. R. Astron. Soc.* **479**, 1711–1721 (2018).
- Malamud, U., Perets, H. B., Schäfer, C. & Burger, C. Collisional formation of massive exomoons of superterrestrial exoplanets. *Mon. Not. R. Astron. Soc.* **492**, 5089–5101 (2020).
- Malamud, U. & Perets, H. B. Tidal disruption of planetary bodies by white dwarfs – I: A hybrid SPH-analytical approach. *Mon. Not. R. Astron. Soc.* **492**, 5561–5581 (2020).
- Malamud, U. & Perets, H. B. Tidal disruption of planetary bodies by white dwarfs – II: Debris disc structure and ejected interstellar asteroids. *Mon. Not. R. Astron. Soc.* **493**, 698–712 (2020).
- Herrmann, W. Constitutive equation for the dynamic compaction of ductile porous materials. *J. Appl. Phys.* **40**, 2490–2499 (1969).
- Carroll, M. & Holt, A. C. Suggested modification of the P – α model for porous material. *J. Appl. Phys.* **43**, 759–761 (1972).
- Jutzi, M., Michel, P., Hiraoka, K., Nakamura, A. M. & Benz, W. Numerical simulations of impacts involving porous bodies. II. Comparison with laboratory experiments. *Icarus* **201**, 802–813 (2009).
- Leleu, A., Jutzi, M. & Rubin, M. The peculiar shapes of Saturn's small inner moons as evidence of mergers of similar-sized moonlets. *Nat. Astron.* **2**, 555–561 (2018).
- Jutzi, M., Benz, W., Toliou, A., Morbidelli, A. & Brasser, R. How primordial is the structure of comet 67P? Combined collisional and dynamical models suggest a late formation. *Astron. Astrophys.* **597**, A61 (2017).
- Rotundi, A. et al. Dust measurements in the coma of comet 67P/Churyumov–Gerasimenko inbound to the sun. *Science* **347**, aac3905 (2015).
- Malamud, U. & Prialnik, D. Modeling Kuiper belt objects Charon, Orcus and Salacia by means of a new equation of state for porous icy bodies. *Icarus* **246**, 21–36 (2015).
- Lorek, S., Gundlach, B., Lacerda, P. & Blum, J. Comet formation in collapsing pebble clouds. What cometary bulk density implies for the cloud mass and dust-to-ice ratio. *Astron. Astrophys.* **587**, A128 (2016).
- Fuller, M. et al. The dust-to-ices ratio in comets and Kuiper belt objects. *Mon. Not. R. Astron. Soc.* **469**, S45–S49 (2017).
- Jutzi, M., Benz, W. & Michel, P. Numerical simulations of impacts involving porous bodies. I. Implementing sub-resolution porosity in a 3D SPH hydrocode. *Icarus* **198**, 242–255 (2008).

58. Grady, E. D. & Kipp, E. Dynamic fracture and fragmentation. In *High-Pressure Shock Compression of Solids* (eds Asay, J. R. & Shahinpoor, M.) 265–322 (Springer, 1993).
59. Benz, W. & Asphaug, E. Impact simulations with fracture. I – Method and tests. *Icarus* **107**, 98 (1994).
60. Benz, W. & Asphaug, E. Catastrophic disruptions revisited. *Icarus* **142**, 5–20 (1999).
61. Jutzi, M. SPH calculations of asteroid disruptions: the role of pressure dependent failure models. *Planet. Space Sci.* **107**, 3–9 (2015).
62. Collins, G. S., Melosh, H. J. & Ivanov, B. A. Modeling damage and deformation in impact simulations. *Meteorit. Planet. Sci.* **39**, 217–231 (2004).

Acknowledgements We acknowledge discussions with D. C. Fabrycky and E. Kite. H.B.P. acknowledges support from the MINERVA Center for Life Under Extreme Planetary Conditions and the Kingsley Fellowship at Caltech. C.M.S. and O.W. acknowledge support by the High Performance and Cloud Computing Group at the Zentrum für Datenverarbeitung of the University of Tübingen, the state of Baden-Württemberg through bwHPC and the German

Research Foundation (DFG) through grant number INST 37/935-1 FUGG. C.M.S. acknowledges support from the DFG through grant number 398488521.

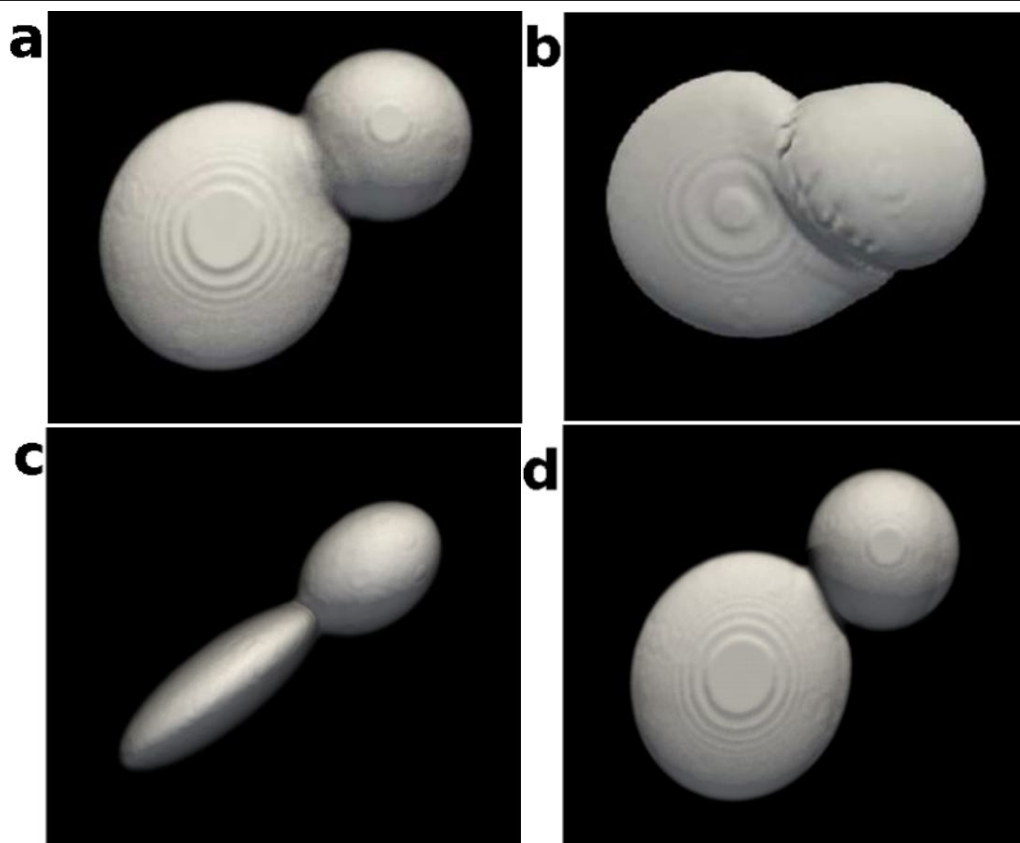
Author contributions E.G. led the project, performed the analytic calculations and ran and analysed the *N*-body simulations. U.M. led the hydrodynamical modelling, its analysis and wrote the hydrodynamical sections. H.B.P. initiated the project and supervised it, suggested the main ideas and concepts and took part in all of the analysis. O.W. ran the hydrodynamical simulations and was the main developer of the porosity module in the hydrodynamical code. C.M.S. developed the hydrodynamical code and supervised the development of the porosity module. E.G. and H.B.P. wrote the main text.

Competing interests The authors declare no competing interests.

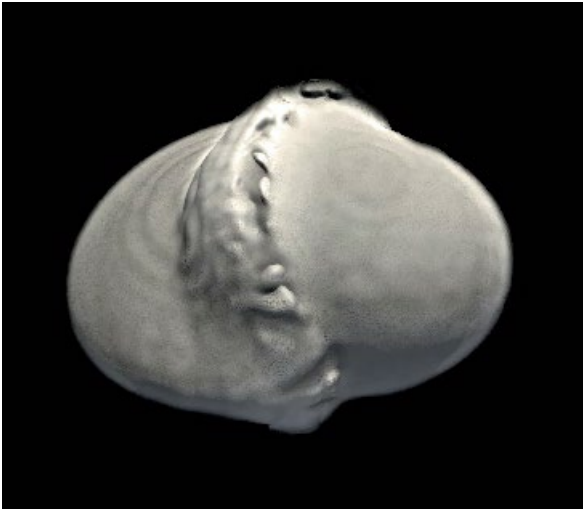
Additional information

Correspondence and requests for materials should be addressed to E.G.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Additional results of the collision models. a, 40° impact angle, medium-strength material. **b**, 40° impact angle, low-strength material. **c, d**, Low-density model (0.5 g cm^{-3}) with an impact angle of 55° and medium-strength material. The edge (**c**) and face (**d**) views are given.



Extended Data Fig. 2 | Additional results of the collision models. 5° impact angle, high-strength material and large escape velocity, $v = 10v_{\text{esc}}$.

Extended Data Table 1 | Merger rate of the binaries in the non-secular regime

	0.2	0.3	0.4	U
N_m	78	99	114	101
N_{80-100}	9	12	18	15
f_i	0.12	0.16	0.18	0.16
f_{80-100}	0.014	0.019	0.029	0.024

N_m , total number of mergers.
 N_{80-100} , number of mergers with inclinations $80^\circ < i < 100^\circ$.
 f_i , total merger fraction normalized to the inclination sampling rate.
 f_{80-100} , merger fraction for only those mergers with inclinations $80^\circ < i < 100^\circ$.

Extended Data Table 2 | Crush curve, plasticity and fragmentation parameters

Type	P_e (Pa)	P_s (Pa)	Y_0 (Pa)	Y_T (Pa)
Low-strength	10^4	10^6	10^4	10^3
Medium-strength	10^5	10^7	10^5	10^4
High-strength	10^6	10^8	10^6	10^5

Observation of topologically enabled unidirectional guided resonances

<https://doi.org/10.1038/s41586-020-2181-4>

Xuefan Yin^{1,2}, Jicheng Jin³, Marin Soljačić², Chao Peng^{1,2✉} & Bo Zhen³

Received: 24 June 2019

Accepted: 22 January 2020

Published online: 22 April 2020

 Check for updates

Unidirectional radiation is important for various optoelectronic applications, such as lasers, grating couplers and optical antennas. However, almost all existing unidirectional emitters rely on the use of materials or structures that forbid outgoing waves—that is, mirrors, which are often bulky, lossy and difficult to fabricate. Here we theoretically propose and experimentally demonstrate a class of resonances in photonic crystal slabs that radiate only towards one side of the slab, with no mirror placed on the other side. These resonances, which we name ‘unidirectional guided resonances’, are found to be topological in nature: they emerge when a pair of half-integer topological charges^{1–3} in the polarization field bounce into each other in momentum space. We experimentally demonstrate unidirectional guided resonances in the telecommunication regime by achieving single-side radiative quality factors as high as 1.6×10^5 . We further demonstrate their topological nature through far-field polarimetry measurements. Our work represents a characteristic example of applying topological principles^{4,5} to control optical fields and could lead to energy-efficient grating couplers and antennas for light detection and ranging.

Topological defects¹, which are characterized by quantized invariants, offer a general description of many exotic phenomena in real space, such as quantum vortices in superfluids and singular optical beams³. It has been recently found that topological defects can also emerge in momentum space, giving rise to interesting effects. One such example is bound states in the continuum⁶ (BICs) in photonic crystal slabs: these guided resonances reside inside the continuous spectrum of extended radiating modes, yet counter-intuitively remain spatially confined and maintain infinitely long lifetimes. Since they were initially proposed⁷, BICs have been demonstrated in a variety of wave systems^{8–20} and have led to various applications^{21,22}. Recently, the topological defect nature of BICs in photonic crystal slabs was discovered: BICs are vortices of polarization major axes in momentum space that carry integer topological charges^{2,23–25}. The lack of a continuous definition of polarization at the vortex centre forbids the emission of far-field radiation from BICs. So far, most BICs have been demonstrated in up–down mirror-symmetric structures^{11,13,26}, in which the observation of no upward radiation necessitates the absence of downward radiation. On the other hand, the existence of unidirectional guided resonances (UGRs)—resonances that radiate only towards one side of a photonic crystal slab, with no mirror placed on the other side—has not been confirmed so far^{27,28}.

Here we theoretically propose and experimentally demonstrate UGRs in photonic crystal slabs that are enabled by topological charges. Specifically, we first split the integer topological charge carried by a BIC into a pair of half-integer topological charges²⁹; each charge corresponds to a circularly polarized resonance. As the structure is continuously varied, the two half-integer topological charges in the downward radiation keep evolving in momentum space until they bounce into each other and, again, act as an integer charge. At this point, downward radiation from

this resonance is disallowed because its far-field polarization is again undefined—this is the topological origin of UGRs. Because up–down mirror symmetry is broken, the UGRs can still radiate towards the top side, unlike traditional BICs⁶, making them potentially useful as low-loss grating couplers to efficiently couple light both on and off chips.

Numerical design and topological interpretation

As a specific example, we consider a one-dimensional periodic photonic crystal slab in which infinitely long bars with gaps of width $w = 358$ nm are defined in a 500-nm-thick silicon layer with refractive index of $n = 3.48$ at a periodicity of $a = 772$ nm (Fig. 1a–d). Both the top and bottom silica cladding layers ($n = 1.46$) are assumed to be semi-infinitely thick. When the sidewalls of the bars are vertical ($\theta = 90^\circ$; Fig. 1b), the photonic crystal slab is up–down and left–right symmetric, and a BIC is found on a transverse electric (TE)-like band (TE1) along the k_x axis off the normal direction at $k_x a / (2\pi) = 0.176$. In this up–down-symmetric structure, the radiative decay rate of a mode towards the top (γ_t ; orange) of the photonic crystal slab is always the same as that towards the bottom (γ_b ; blue), and both rates are reduced to 0 at the BIC (middle panels, Fig. 1b). Fundamentally, this BIC is a topological defect in the far-field polarization major axes that carries an integer topological charge of $q = 1$, defined as:

$$q = \oint_C d\mathbf{k} \cdot \nabla_{\mathbf{k}} \phi(\mathbf{k}) \quad (1)$$

Here $\phi(\mathbf{k})$ is the angle between polarization major axis and the x axis and \mathbf{k} is the in-plane wave vector. C is a yellow closed path in Fig. 1b, which goes around the BIC in the counter-clockwise direction.

¹State Key Laboratory of Advanced Optical Communication Systems and Networks, Department of Electronics and Frontiers Science Center for Nano-optoelectronics, Peking University, Beijing, China. ²Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA. ³Department of Physics and Astronomy, University of Pennsylvania, Philadelphia, PA, USA. ✉e-mail: pengchao@pku.edu.cn

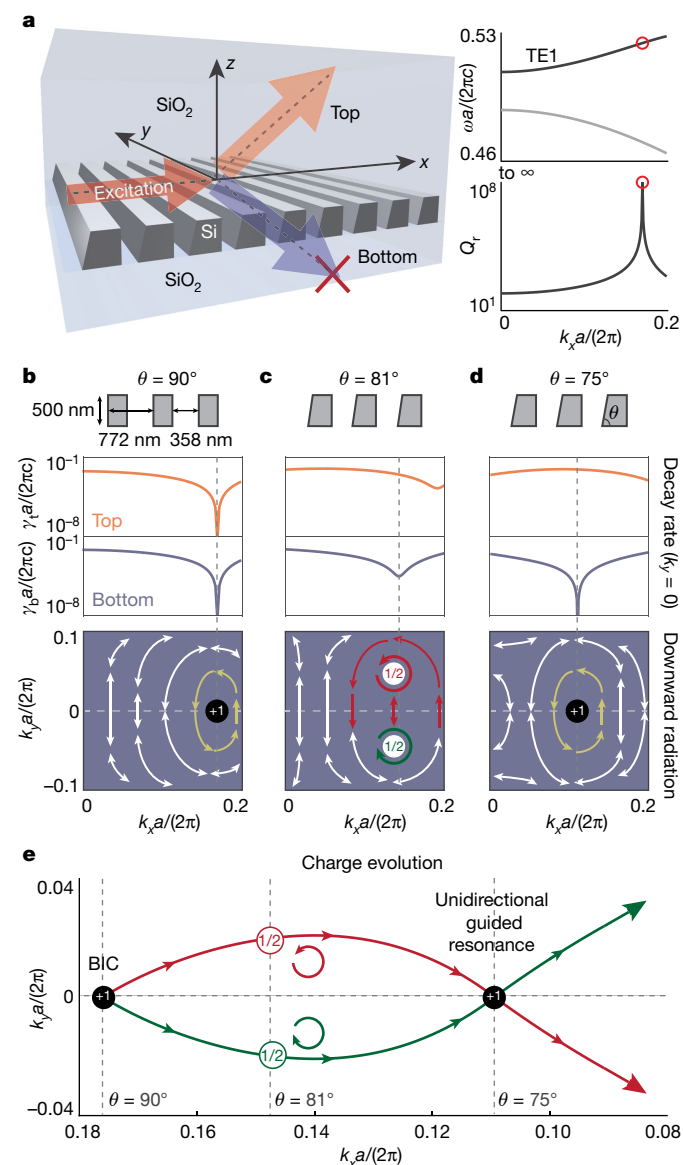


Fig. 1 | UGRs and their topological nature. **a**, Schematic of a photonic crystal slab supporting a unidirectional guided resonance (UGR). When the sidewall is vertical, a BIC is found on a TE band, labelled by red circles in the right panels. Q_t accounts for radiative loss towards both top and bottom, $Q_t = \omega/(\gamma_t + \gamma_b)$. **b**, When the sidewall is vertical, radiative losses from resonances towards the top (γ_t ; orange line) are equal to those towards the bottom (γ_b ; blue line) owing to up-down symmetry, and both reduce to 0 at the BIC. The polarization major axes wind around the BIC and have a topological charge of $q = 1$. **c**, When the sidewall is tilted from the vertical direction, the $q = 1$ topological charge splits into a pair of half charges ($q = 1/2$) with opposite helicities, LCP (red) and RCP (green). **d**, When the sidewall angle θ changes to 75° , a UGR is achieved: radiation towards the bottom is eliminated ($\gamma_b = 0$) while radiation emitted towards the top (γ_t) remains finite. **e**, Trajectories traced by the two half charges (red and green) in momentum space as θ is varied.

When one of the sidewalls is tilted away from the vertical direction ($\theta = 81^\circ$; Fig. 1c), the photonic crystal slab is no longer up-down symmetric, and γ_t and γ_b are no longer simply related. No BIC exists in this structure any more; the radiative decay rate towards the top or the bottom ($\gamma_{t,b}$) never reaches 0 (middle panel). On the other hand, the total winding of the polarization major axes remains $+2\pi$ because the winding number is a conserved quantity. Consequently, the integer charge $q = 1$ is split into two half-integer charges of $q = 1/2$, each being

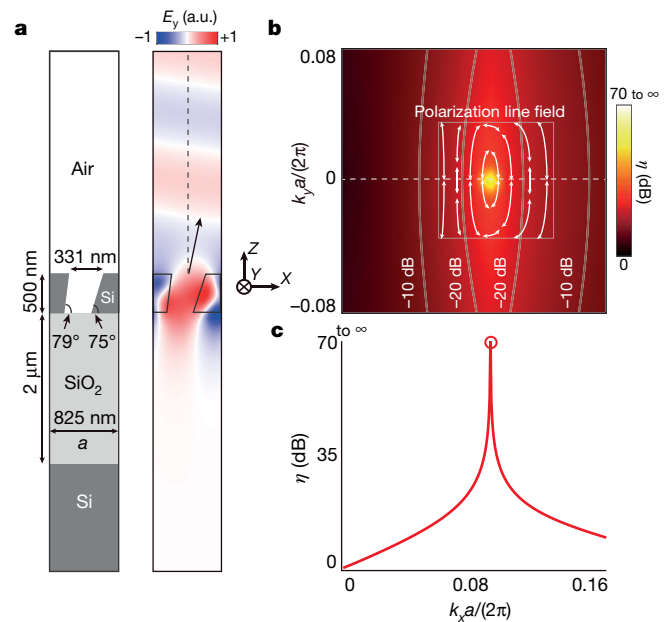


Fig. 2 | Numerical confirmation of a UGR. **a**, Unit cell design of a photonic crystal slab supporting a UGR that only radiates towards the top, as shown in its mode profile (E_y). a.u., arbitrary units. **b**, **c**, Colour map (**b**) and line cut (**c**) showing that the asymmetry ratio between upward and downward radiative loss, $\eta = \gamma_t/\gamma_b$, diverges to infinity at the UGR. The polarization major axes (arrows) show a $+2\pi$ winding around the UGR, which is consistent with Fig. 1d.

a circularly polarized resonance (bottom panel). The two half-integer charges are related to each other by the y -mirror symmetry of the structure, which also guarantees that these two circularly polarized resonances are opposite in helicity: left-handed circularly polarized (LCP) for one (red) and right-handed circularly polarized (RCP) for the other (green).

When the sidewall is further tilted, the two half-integer charges in the downward radiation keep moving in momentum space, following the trajectories shown in Fig. 1e: red for LCP and green for RCP. Neither of the radiative decay rates ($\gamma_{t,b}$) is reduced to 0 until θ is decreased to 75° (Fig. 1d), where the LCP and RCP trajectories meet on the k_x axis. At this point, any downward radiation needs to be both LCP and RCP at the same time, which can never be satisfied. As a result, this guided resonance cannot have any downward radiation, even without a mirror on the bottom—this is what we call a UGR. From the viewpoint of topology, UGRs can be understood as the merging point between two half-integer charges, where they act like an integer charge, forbidding any radiative loss. This topological interpretation agrees with our numerical simulation results, where γ_b reaches 0 whereas γ_t remains finite (middle panel of Fig. 1d). We note that the lack of certain symmetries in our structure (both C_2 and up-down mirror) is crucial to achieve UGRs; see Supplementary Information sections 1–3 for more details.

Next, we present our UGR design (Fig. 2a). The photonic crystal slab consists of a periodic array of one-dimensional bars defined in a 500-nm-thick silicon-on-insulator wafer at a periodicity of $a = 825$ nm (left panel). The top cladding material is air and the bottom cladding is SiO_2 . The sidewalls are tilted to specific angles, $\theta_l = 79^\circ$ and $\theta_r = 75^\circ$, to achieve a UGR: as shown in the E_y mode profile (right panel), the downward radiation γ_b is considerably lower, by more than 70 dB, than the upward radiation γ_t . The asymmetry ratio between upward and downward radiation intensity, $\eta = \gamma_t/\gamma_b$, is calculated for different k points (colour map, Fig. 2b), where the extremely bright spot marks the location of the UGR at $k_x a/(2\pi) = 0.0854$. A line-cut of the colour map along the k_x axis shows the asymmetry ratio η diverging into infinity, which is the characteristic feature of unidirectional radiation (Fig. 2c).

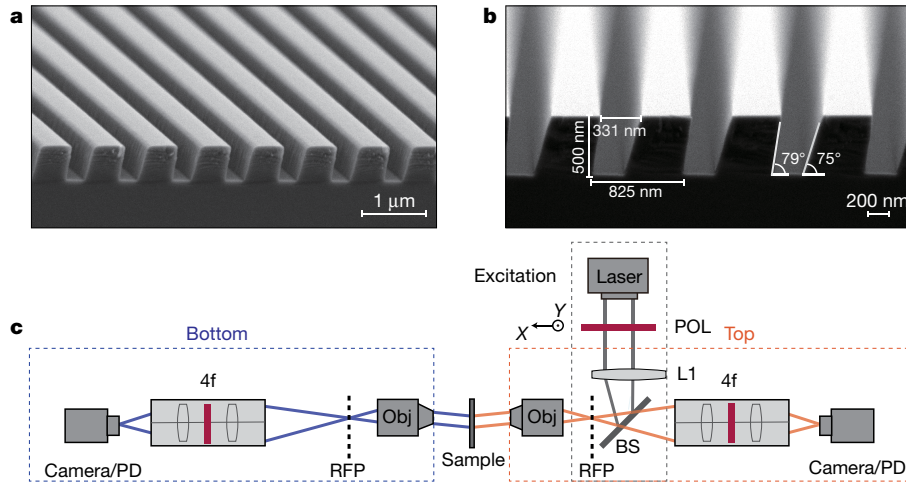


Fig. 3 | Fabricated sample and experimental setup. **a, b** Scanning electron microscope images of the fabricated photonic crystal sample from tilted (a) and side (b) views. **c**, Schematic of the setup used to independently measure

the upward and downward radiation intensities from the guided resonances in the photonic crystal sample. L, lens; Obj, objective; RFP, rear focal plane; PD, photodetector; POL, polarizer; BS, beam splitter; 4f, relay 4f optical system.

Overlaid on the colour map in Fig. 2b is the plot of the polarization major axes for the downward far-field radiation from nearby resonances. An integer winding of the polarization major axes, $q = 1$, is observed around the UGR, which is consistent with the topological interpretation presented in Fig. 1d.

Sample fabrication and experimental setup

To verify our theoretical findings, we fabricate photonic crystal samples with UGRs using plasma-enhanced chemical vapour deposition, electron-beam lithography and reactive ion etching (RIE) processes. The scanning electron microscope images are shown in Fig. 3a, b. Briefly, a thermal SiO₂ layer with a thickness of approximately 110 nm is first deposited on the wafer as the hard mask. Unlike standard RIE processes that use horizontal substrates, our sample is placed on a wedged substrate that allows us to etch the silicon layer at a slanted angle; as a result, high-quality air gaps with tilted sidewalls are achieved (Fig. 3b). Because of the shadowing effect, the angles of the left and right sidewalls are not identical: $\theta_l = 79^\circ$ and $\theta_r = 75^\circ$. The width of the air gap, w , is swept from 320 nm to 340 nm to best capture the UGR design at $w = 331$ nm. See Methods for more details about the fabrication.

To demonstrate UGRs, the upward and downward radiative decay rates from our fabricated samples are independently characterized using the experimental setup schematically shown in Fig. 3c. A tunable telecommunication laser in the C+L band is first sent through a polarizer in the x direction before it is focused by a lens (L1) onto the rear focal plane of an infinity-corrected objective lens. To achieve on-resonance coupling, the incident angle is tuned for each excitation wavelength λ by moving L1 in the x - y plane, exciting a resonance in the sample. Each excited resonance radiates towards the top (bottom) according to its radiative decay rate γ_t (γ_b) into this channel. Upward (downward) far-field radiation from this resonance is then collected by the confocal setup shown on the right (left), marked with an orange (blue) dashed box, where the beam is shrunk by 0.67 times through a 4f system to best fit the camera. This on-resonance excitation scheme is similar to previously reported results^{30,31}. See Methods for more details on the experimental setup.

Experimental results

As an example, the experimental comparison between upward and downward radiation from a resonance at $\lambda = 1,551$ nm is shown in Fig. 4a.

Here, the excitation laser is on resonance with a mode on the k_y axis at $k_y a / (2\pi) = 0.01$. Momentum space is labelled with respect to the known numerical aperture of the objectives (NA = 0.26), shown as white circles. The characteristic feature of the UGR—marked by a white arrow on the k_x axis—is qualitatively shown in the comparison between the two figures: for resonances near the white arrow, the downward radiation (X', Y', Z') is always much weaker than the upward radiation (X, Y, Z). On the other hand, for resonances far from the UGRs (for example, to the left of the k_y axis), the upward and downward radiation are comparable. We note that although UGRs radiate only towards a single side (top), their in-plane propagation is not immune to back-scattering from fabrication disorder such as the chiral edge states in a Chern insulator, because our structure is reciprocal.

A more quantitative demonstration of the UGRs is achieved by measuring the up-down asymmetry ratio $\eta = \gamma_t / \gamma_b$ of the resonances. Two movable pinholes (not shown in Fig. 3c) with diameters of 300 μm are placed at the image planes of the rear focal planes of the objectives to select specific k points. Three examples are shown in Fig. 4b, where upward (X, Y, Z) and downward (X', Y', Z') radiation intensities are measured by two photodetectors as the excitation wavelength scans through the three resonances. As expected, all measured spectra exhibit symmetric Lorentzian features³⁰: the excitation efficiency reaches its maximum when the excitation is on resonance, which happens at $\lambda = 1,553.7$ nm, 1,551.2 nm and 1,549.4 nm. Accordingly, both the central wavelengths and the total quality factors of the resonances can be extracted by fitting the experimental results. By repeating this procedure for all resonances along the X - Z line, we achieve good agreement between experiments (red crosses) and numerical simulation (blue line, Fig. 4c).

We further measure the downward radiative decay rate of the resonances, $\gamma_b = \omega / Q_b$, and show that it is reduced to 0 at the UGR. Here, ω is the resonance frequency and Q_b is the radiative quality factor that accounts only for the downward radiation. In practice, the observed total loss $\gamma_{\text{tot}} = \omega / Q_{\text{tot}}$ is composed of non-radiative loss $\gamma_{\text{non-rad}} = \omega / Q_{\text{non-rad}}$ (including absorption, scattering, and lateral leakage), as well as radiative losses towards the top and bottom:

$$\gamma_{\text{tot}} = \gamma_{\text{non-rad}} + \gamma_t + \gamma_b \quad (2)$$

Because these resonances are close in momentum space and share similar mode profiles, it is reasonable to assume that they share a similar non-radiative quality factor, which is found to be $Q_{\text{non-rad}} = 2,080$ through

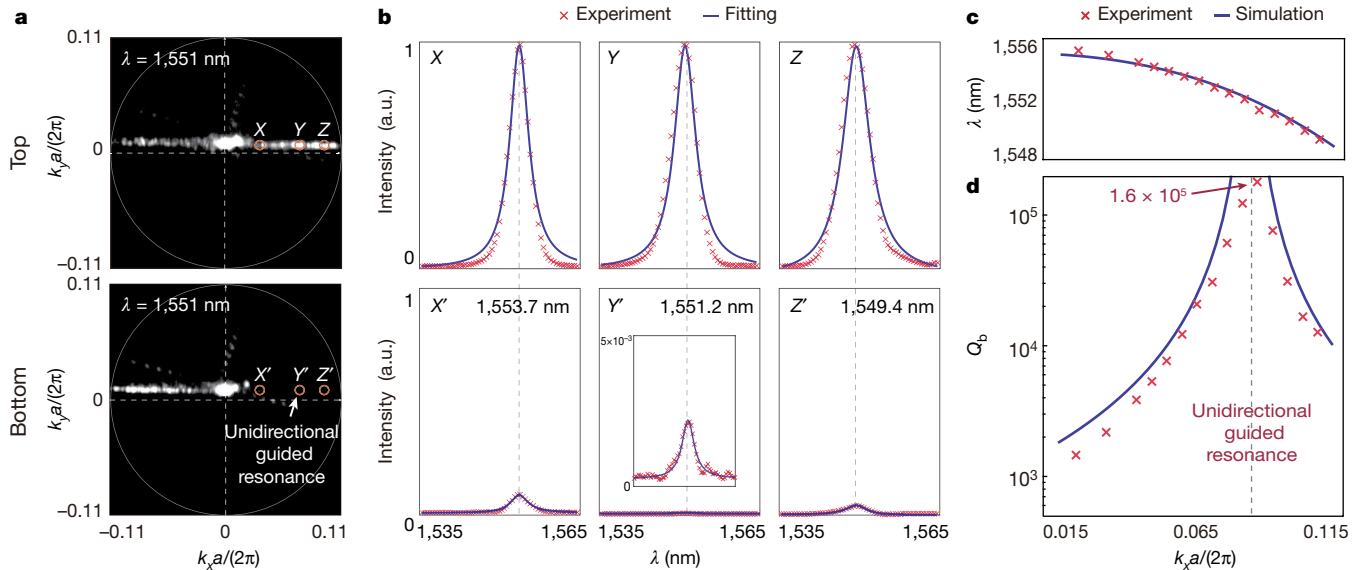


Fig. 4 | Observation of UGRs. **a**, Upward and downward radiation intensities from resonances under an excitation wavelength of 1,551 nm. In the vicinity of the UGR on the k_x axis, marked by a white arrow, the downward radiation intensities (X' , Y' , Z') are considerably suppressed compared to the upward

radiation (X , Y , Z). **b**, Upward and downward radiation intensities from the resonances as the excitation wavelength scans from 1,535 nm to 1,565 nm. **c**, **d**, Experimental results (red crosses) of the band structure (**c**) and Q_b (**d**), showing good agreement with the simulation results (blue lines).

numerical fitting (see Methods for details). Upward and downward radiative decay rates can be further separated based on the measured asymmetry ratio $\eta = \gamma_u/\gamma_b$ (see Extended Data Fig. 3 for measurement results of η). Experimentally extracted Q_b values are presented in Fig. 4c as red crosses, showing good agreement with the numerical simulation results (blue line). In particular, the fact that the bottom radiation γ_b is reduced to almost 0 at $k_x a/(2\pi) = 0.088$ proves the existence of UGR.

To demonstrate the topological nature of UGRs, we perform polarimetry measurements³¹ on a series of five samples with slightly different widths w . For each sample, we experimentally locate the two half-integer charges in momentum space (symbols in Fig. 5a), which show good agreement with the simulation results (dashed lines). See Methods for more experimental details of the polarimetry measurements. The perfect design with a UGR is marked with an arrow. As shown, when w increases from 0.399 a (marked by inverted triangles) to 0.403 a (diamonds), the two half-integer charges switch positions.

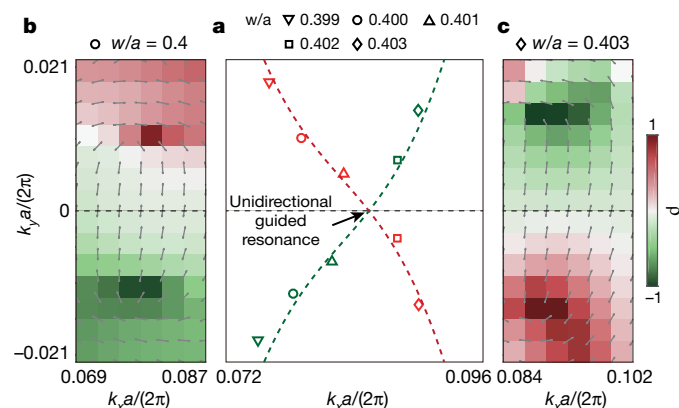


Fig. 5 | Observation of the topological nature of UGRs. **a**, Experimentally measured trajectories of half-integer charges from five samples with different widths w (symbols), showing good agreement with simulation results (dashed lines). **b**, **c**, The LCP and RCP resonances switch positions as w increases from 0.4 a (**b**) to 0.403 a (**c**), with the merging point being the UGR. The colour map shows the measured ellipticity of the downward radiation fields.

This switching behaviour is further confirmed by measuring the ellipticity ρ of the far-field polarization: when w increases from 0.4 a (Fig. 5b) to 0.403 a (Fig. 5c), the LCP (RCP) resonance, shown in red (green), moves from the top (bottom) to the bottom (top) half of the momentum space. Taken together, these experimental results confirm our topological interpretation shown in Fig. 1d: UGRs arise when two half-integer charges with opposite helicities bounce into each other in momentum space.

To summarize, we present a type of resonance, which we call UGR, that radiates only towards the top of a photonic crystal slab, even without a bottom mirror. We experimentally demonstrate their existence by showing that the downward radiation field vanishes. Through polarimetry measurements, we further demonstrate the topological nature of these resonances as the merging point between half-integer topological charges. Owing to their unique properties, UGRs could be used as energy-efficient grating couplers (see Methods for discussion) with further applications in photonic-crystal surface-emitting lasers, light detection and ranging antennas.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2181-4>.

- Mermin, N. D. The topological theory of defects in ordered media. *Rev. Mod. Phys.* **51**, 591–648 (1979).
- Zhen, B., Hsu, C. W., Lu, L., Stone, A. D. & Soljačić, M. Topological nature of optical bound states in the continuum. *Phys. Rev. Lett.* **113**, 257401 (2014).
- Gbur, G. J. *Singular Optics* (CRC Press, 2016).
- Lu, L., Joannopoulos, J. D. & Soljačić, M. Topological photonics. *Nat. Photon.* **8**, 821–829 (2014).
- Ozawa, T. et al. Topological photonics. *Rev. Mod. Phys.* **91**, 015006 (2019).
- Hsu, C. W., Zhen, B., Stone, A. D., Joannopoulos, J. D. & Soljačić, M. Bound states in the continuum. *Nat. Rev. Mater.* **1**, 16048 (2016).
- von Neuman, J. & Wigner, E. Über merkwürdige diskrete Eigenwerte. Über das Verhalten von Eigenwerten bei adiabatischen Prozessen. *Phys. Z.* **30**, 467–470 (1929).
- Friedrich, H. & Wintgen, D. Interfering resonances and bound states in the continuum. *Phys. Rev. A* **32**, 3231–3242 (1985).

9. Fan, S. & Joannopoulos, J. D. Analysis of guided resonances in photonic crystal slabs. *Phys. Rev. B* **65**, 235112 (2002).
10. Plotnik, Y. et al. Experimental observation of optical bound states in the continuum. *Phys. Rev. Lett.* **107**, 183901 (2011).
11. Hsu, C. W. et al. Observation of trapped light within the radiation continuum. *Nature* **499**, 188–191 (2013).
12. Corielli, G., Della Valle, G., Crespi, A., Osellame, R. & Longhi, S. Observation of surface states with algebraic localization. *Phys. Rev. Lett.* **111**, 220403 (2013).
13. Kodigala, A. et al. Lasing action from photonic bound states in continuum. *Nature* **541**, 196–199 (2017).
14. Gomis-Bresco, J., Artigas, D. & Torner, L. Anisotropy-induced photonic bound states in the continuum. *Nat. Photon.* **11**, 232–236 (2017).
15. Molina, M. I., Miroshnichenko, A. E. & Kivshar, Y. S. Surface bound states in the continuum. *Phys. Rev. Lett.* **108**, 070401 (2012).
16. Carletti, L., Koshelev, K., De Angelis, C. & Kivshar, Y. Giant nonlinear response at the nanoscale driven by bound states in the continuum. *Phys. Rev. Lett.* **121**, 033903 (2018).
17. Monticone, F. & Alù, A. Embedded photonic eigenvalues in 3D nanostructures. *Phys. Rev. Lett.* **112**, 213903 (2014).
18. Liu, Z. et al. High-Q quasibound states in the continuum for nonlinear metasurfaces. *Phys. Rev. Lett.* **123**, 253901 (2019).
19. Lim, T. C. & Farnell, G. W. Character of pseudo surface waves on anisotropic crystals. *J. Acoust. Soc. Am.* **45**, 845–851 (1969).
20. Cobelli, P. J., Pagneux, V., Maurel, A. & Petitjeans, P. Experimental observation of trapped modes in a water wave channel. *Europhys. Lett.* **88**, 20006 (2009).
21. Hirose, K. et al. Watt-class high-power, high-beam-quality photonic-crystal lasers. *Nat. Photon.* **8**, 406 (2014).
22. Chow, E., Grot, A., Mirkarimi, L. W., Sigalas, M. & Girolami, G. Ultracompact biochemical sensor built with two-dimensional photonic crystal microcavity. *Opt. Lett.* **29**, 1093–1095 (2004).
23. Bulgakov, E. N. & Maksimov, D. N. Topological bound states in the continuum in arrays of dielectric spheres. *Phys. Rev. Lett.* **118**, 267401 (2017).
24. Zhang, Y. et al. Observation of polarization vortices in momentum space. *Phys. Rev. Lett.* **120**, 186103 (2018).
25. Doeleman, H. M., Monticone, F., den Hollander, W., Andrea, A. & Koenderink, A. F. Experimental observation of a polarization vortex at an optical bound state in the continuum. *Nat. Photon.* **12**, 397–401 (2018).
26. Yang, Y., Peng, C., Liang, Y., Li, Z. & Noda, S. Analytical perspective for bound states in the continuum in photonic crystal slabs. *Phys. Rev. Lett.* **113**, 037401 (2014).
27. Zhou, H. et al. Perfect single-sided radiation and absorption without mirrors. *Optica* **3**, 1079–1086 (2016).
28. Wang, K. X., Yu, Z., Sandhu, S. & Fan, S. Fundamental bounds on decay rates in asymmetric single-mode optical resonators. *Opt. Lett.* **38**, 100–102 (2013).
29. Liu, W. et al. Circularly polarized states spawning from bound states in the continuum. *Phys. Rev. Lett.* **123**, 116104 (2019).
30. Jin, J. et al. Topologically enabled ultrahigh-Q guided resonances robust to out-of-plane scattering. *Nature* **574**, 501–504 (2019).
31. Zhou, H. et al. Observation of bulk Fermi arc and polarization half charge from paired exceptional points. *Science* **359**, 1009–1012 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Numerical simulation details

All simulations are performed using the COMSOL Multiphysics 5.2 'Radio Frequency' module in the frequency domain. Two-dimensional models (in the x - y plane) are created to simulate one-dimensional photonic crystal slabs with perfectly matching layers along the y direction. Periodic (Floquet) boundary conditions are applied in the x direction. The meshing resolution is set to adequately capture resonances with Q values of up to 10^9 . The eigenvalue solver is used to calculate the band structures and quality factors. To calculate the asymmetric radiation ratios η , two surface probes—one above and one below the structure—are used to calculate the radiative fields towards both sides.

Sample fabrication

The sample is fabricated on a single-side polished silicon-on-insulator wafer with a silicon layer thickness of 500 nm, a silica layer of 2 μm and a silicon substrate of $\sim 725 \mu\text{m}$. The step-by-step fabrication process is illustrated in Extended Data Fig. 1a. The wafer is first cleaved into 1 cm \times 1 cm chips and cleaned. The photonic crystal pattern is then defined using electron-beam lithography. A 110-nm-thick SiO_2 film is thermally deposited using plasma-enhanced chemical vapour deposition as the hard mask. A layer of ZEP520A photo-resist (340 nm thick) is spin-coated on the SiO_2 layer. The photonic crystal patterns are defined in the photo-resist layer (Elionix FLS-125), which is then developed using o-xylene (>95%). The chip is then placed on a customized wedged holder made of Al_2O_3 with a slanted angle of 26° for RIE. The photonic crystal patterns are first transferred onto the SiO_2 hard mask using CF_4 , and are then transferred onto Si using Cl_2 gas. The residue hard mask is removed using BHF wet etching. Finally, the bottom side is treated with chemical-mechanical polishing for the measurements of the bottom radiation fields. The size of the sample is 500 μm \times 500 μm . Owing to the shadowing effect, the tilt angles of the two sidewalls are not identical, as illustrated in Extended Data Fig. 1b. To best capture the UGR design at $w = 331 \text{ nm}$, the width of the air gap is varied from 320 nm to 340 nm.

Measurement of asymmetry ratio η and single-side quality factors

As illustrated in Extended Data Fig. 2, a x -polarized tunable telecommunication laser (Santec TSL-550, C+L band) is sent through a chopper for lock-in detection and is focused by a lens (L1) onto the rear focal plane of the objective to define the excitation angle. Two identical arms are used to measure the two radiation fields from the top and the bottom. In each arm, a two-stage 4f system is used to adjust the magnification ratio. After passing through an orthogonal polarizer in the y direction, the radiation field is collected using a photodetector and a camera. To measure radiation fields from resonances from a specific k point, two movable pin holes with diameters of 300 μm are placed at the Fourier planes to select the desired k point. Upward and downward radiation fields go through two identical pin holes and are then measured using two identical photodetectors (PDA10DT-EC). Each photodetector is connected to a lock-in amplifier (SRS SR830). A flip mirror is used to switch between the camera that images the light-scattering patterns and the photodetector.

A 'cross-polarization filtering' technique is used to suppress unwanted reflections, similarly to some previous works^{32,33}. Specifically, unwanted reflections (caused by lenses or other optical surfaces) mostly maintain the incident polarization, whereas most radiation fields from guided resonances do not. By placing two orthogonal polarizers in the optical path along the x axis (for excitation) and y axis (for observation), unwanted reflection is greatly suppressed. This setup also transforms typical asymmetric Fano lineshapes into nearly symmetric Lorentzian lineshapes. An 'on-resonance pumping' technique is also used in the setup, similarly to previous works³⁴. As the photonic

crystal structure shows little dispersion along k_y but strong dispersion along k_x , the scattering patterns are almost straight lines parallel to the x axis.

The central wavelengths (Fig. 4c) and total quality factors (Q_{tot} ; Extended Data Fig. 3a) of the guided resonances are extracted by numerically fitting the measured spectra; examples are shown in Fig. 4b. As both upward and downward radiation fields are measured in our setup, the ratio between upward and downward decay rate, η , is achieved directly. The observed total quality factor Q_{tot} includes contributions from: (1) non-radiative losses due to material absorption, scattering from surface roughness and in-plane lateral leakage and (2) radiative losses due to upward and downward radiation. This relationship can be written as:

$$\frac{1}{Q_{\text{tot}}} = \frac{1}{Q_{\text{non-rad}}} + \frac{1}{Q_r} \quad (3)$$

The radiative losses can be further separated into top and bottom channels: $1/Q_r = 1/Q_t + 1/Q_b$. As shown in Extended Data Fig. 3a, by comparing the measured Q_{tot} (blue crosses) to the calculated radiative quality factors from an ideal unidirectional design (red line), $Q_{\text{non-rad}} = 2,080$ is extracted. By design, $Q_{\text{non-rad}}$ is much larger than Q_r (200 to 600), so the energy loss is dominated by radiation. Furthermore, the simulation results from disordered designs (green circles) are presented as a reference, where the air-gap locations and widths fluctuate with a standard deviation of 1 nm. Using the measured asymmetry ratio $\eta = \gamma_t/\gamma_b = Q_b/Q_t$, equation (3) can be written as

$$\frac{1}{Q_{\text{tot}}} = \frac{1}{Q_{\text{non-rad}}} + \frac{\eta + 1}{Q_b} \quad (4)$$

Using this relationship, the single-side quality factors Q_b can be calculated accordingly.

Polarimetry measurement setup

To demonstrate the topological nature of UGRs, we perform polarimetry measurements on the downward radiation fields. The experimental setup is schematically shown in Extended Data Fig. 4. Unlike the previous setup, which uses a continuous-wave tunable laser, this setup uses a broadband amplified spontaneous emission light source with a centre wavelength of 1,550 nm, a bandwidth of 40 nm and an output power of 10 dB m. The incident light excites the sample along the k_y axis, and the incident angle is varied between -1.3° and 1.3° at a step size of 0.3° , which is controlled by lens L1. Owing to the broad bandwidth of the excitation, all resonances at a given incident angle are excited.

A standard polarimetry measurement is then performed on the scattered light to determine the polarization state of each resonance. Specifically, the scattered light intensity is measured after passing through six configurations of a polarizer and a quarter-wave plate (QWP): (1) no QWP, polarizer oriented along the x axis; (2) no QWP, polarizer along the y axis; (3) no QWP, polarizer at 45° with respect to the x axis; (4) no QWP, polarizer at 135° with respect to the x axis; (5) QWP fast axis at 45° with respect to the x axis, polarizer along the y axis; (6) QWP fast axis at 135° with respect to the x axis, polarizer along the y axis. This set of measurements allows us to fully reconstruct the polarization state of each resonance³⁵ through the Stokes parameters:

$$\begin{aligned} S_0 &= |E_x|^2 + |E_y|^2 \\ S_1 &= |E_y|^2 - |E_x|^2 \\ S_2 &= 2|E_y E_x| \cos(\Delta\delta) \\ S_3 &= 2|E_y E_x| \sin(\Delta\delta) \end{aligned} \quad (5)$$

Here, $\mathbf{E} = E_x e^{i\omega t} \hat{\mathbf{e}}_x + E_y e^{i\omega t + \Delta\delta} \hat{\mathbf{e}}_y$. Specifically, the ellipticity $\rho = S_y/S_x$ is maximized (+1) or minimized (−1) when $|E_x| = |E_y|$ and $\Delta\delta = \pm\pi/2$, which correspond to the LCP and RCP resonances, respectively. This allows us to locate the half-integer topological charges $q = \pm 1/2$ in momentum space by measuring the maximum and minimum ellipticity ρ of the scattered light.

The ellipticity measurement results for five samples are shown in Extended Data Fig. 5. All samples share the same design, except for the air-gap width w , which varies between $0.399a$ (Extended Data Fig. 5a) and $0.403a$ (Extended Data Fig. 5e). As w/a is varied, the two half-integer charges (with opposite ellipticities) approach and bounce into each other before they move apart. The transition point corresponds to the UGR design, which is confirmed by the switching of the ellipticity before and after the transition; namely, LCP (RCP) is initially in the top (bottom) half plane, as in Extended Data Fig. 5a, and moves to the bottom (top) at the end, as in Extended Data Fig. 5e. These experimental results are in good agreement with the simulation results (Supplementary Fig. 4) and with our topological interpretation presented in Fig. 1e.

Robustness of the UGRs to fabrication errors

In practice, fabricated samples inevitably deviate from their designs because of fabrication errors or imperfections. Here we analyse the factors limiting the performance of UGRs in realistic samples. The periodicity of photonic crystal is limited by the accuracy of the electron-beam lithography; however, this is often not the limiting factor. In comparison, it is more challenging to fabricate the air gaps (both width and tilt angles) exactly as designed, owing to the accuracy of the etching processes. First, we assume that the fabricated sample deviates steadily from the ideal design ($a = 825$ nm, $w = 352$ nm, $\theta_L = 79^\circ$, $\theta_R = 75^\circ$) in terms of (1) air-gap width, $\Delta w = \pm 2.5$ nm and (2) sidewall angle, $\Delta\theta = \pm 1^\circ$. The simulation results in Extended Data Fig. 6a, b confirm that when the parameters are slightly different from the ideal design, the asymmetry ratio remains high, as expected. Furthermore, owing to the topological nature of UGRs, a fixed deviation in one parameter can be compensated by another parameter to restore the perfect elimination of downward radiation fields. For example, as shown in Extended Data Fig. 6c, a change of $\Delta\theta = -1^\circ$ in the etching angle can be compensated by changing the air-gap width from $w = 352$ nm to $w = 365$ nm, where the UGR is restored.

Meanwhile, random fluctuations are also inevitable in fabricated samples and they induce scattering losses and lower the asymmetry ratios. In this part of the analysis, we assume that the tilted angles are fixed while the air-gap locations and widths fluctuate randomly from the ideal design with a standard deviation of 1 nm, which is estimated from the scanning electron microscope images. The average Q_{tot} and asymmetry ratios for disordered samples are obtained from simulations, which are compared with the ideal design and the experimental results, as shown in Extended Data Fig. 3b. Q_{tot} drops owing to scattering losses. The asymmetry ratio is reduced to approximately 50 dB at its peak (from 70 dB) but remains higher than 35 dB in the vicinity of the UGR, demonstrating that our design is robust to fabrication errors and uncertainties. The ‘cross-polarization filtering’ technique also allows us to measure the asymmetry ratio for any k points. As confirmed by simulation and experimental results (Extended Data Fig. 7), the asymmetry ratios remain higher than 35 dB as the excitation deviates from $k_y = 0$ to $k_y a/(2\pi) = 0.06$. This provides a 6° tolerance in the polishing angle of the angled fibre couplers, which is reasonable in practice.

Prospects of using UGRs as grating couplers

Highly directional radiation is desirable in on-chip optoelectronic devices such as lasers, LIDAR antennas and grating couplers. Although grating couplers having been studied extensively, their performances are still not optimal, with one major challenge arising from unwanted

downward radiation losses towards the handle wafer side^{36,37}. Several mechanisms have been proposed to achieve highly directional radiation, including non-resonant blazed gratings and resonance-based dual-layer gratings. Some relevant works^{36,38–44} are listed in Extended Data Table 1 for a comparison with our work, which is based on topology. The measured asymmetry ratio reaches a maximum of 27.7 dB; namely, 99.8% of the radiation field is upward and 0.2% is downward (Extended Data Fig. 8). Near the UGR, strong suppression of the downward radiation is achieved across a reasonably broad bandwidth: over 90% of the upward radiation energy is maintained within a 26 nm bandwidth from 1,536 nm to 1,562 nm, as shown in Extended Data Fig. 8a. Furthermore, we achieve robust suppression of downward radiation at different out-coupling angles between 5° and 11° , as shown in Extended Data Fig. 8b. Although we have not fully characterized the fibre-to-waveguide losses for our design, the UGRs that we demonstrate here naturally eliminate downward radiation and provide a practical and effective method to suppress downward radiative losses.

Data availability

The datasets generated and analysed during the current study are available from the corresponding author upon request.

32. Wang, Z. et al. Mode splitting in high-index-contrast grating with mini-scale finite size. *Opt. Lett.* **41**, 3872–3875 (2016).
33. Lv, J. et al. Demonstration of a thermo-optic phase shifter by utilizing high-Q resonance in high-index-contrast grating. *Opt. Lett.* **43**, 827–830 (2018).
34. Regan, E. C. et al. Direct imaging of isofrequency contours in photonic structures. *Sci. Adv.* **2**, e1601591 (2016).
35. McMaster, W. H. Polarization and the stokes parameters. *Am. J. Phys.* **22**, 351–362 (1954).
36. Notaros, J. et al. Ultra-efficient CMOS fiber-to-chip grating couplers. In *Optical Fiber Communications Conf. Exhib.* 1–3 (IEEE, 2016).
37. Wade, M. T. et al. 75% efficient wide bandwidth grating couplers in a 45 nm microelectronics cmos process. In *IEEE Optical Interconnects Conf.* 46–47 (IEEE, 2015).
38. Wang, B., Jiang, J. & Nordin, G. P. Compact slanted grating couplers. *Opt. Express* **12**, 3313–3326 (2004).
39. Li, M. & Sheard, S. J. Waveguide couplers using parallelogramic-shaped blazed gratings. *Opt. Commun.* **109**, 239–245 (1994).
40. Hagberg, M., Eriksson, N. & Larsson, A. Investigation of high-efficiency surface-emitting lasers with blazed grating outcouplers. *IEEE J. Quantum Electron.* **32**, 1596–1605 (1996).
41. Eriksson, N., Hagberg, M. & Larsson, A. Highly directional grating outcouplers with tailorable radiation characteristics. *IEEE J. Quantum Electron.* **32**, 1038–1047 (1996).
42. Notaros, J. & Popovi, M. A. Band-structure approach to synthesis of grating couplers with ultra-high coupling efficiency and directivity. In *Optical Fiber Communications Conf. Exhib.* 1–3 (IEEE, 2015).
43. Michaels, A. & Yablonovitch, E. Inverse design of near unity efficiency perfectly vertical grating couplers. *Opt. Express* **26**, 4766–4779 (2018).
44. Dai, M. et al. Highly efficient and perfectly vertical chip-to-fiber dual-layer grating coupler. *Opt. Express* **23**, 1691–1698 (2015).

Acknowledgements We thank L. He for discussion, V. Yoshioka for reading the manuscript and Z. Zhang for helping to conduct the experiments. C.P. was supported by the National Natural Science Foundation of China under grant number 61922004. J.J. and B.Z. were sponsored by the US Army Research Office under grant number W911NF-19-1-0087. The simulations were supported by the High-performance Computing Platform of Peking University. The project was partially supported by AFRL contract FA8650-16-D-5403 and MIT Lincoln Laboratory contract 7000371273, as well as by the Army Research Office, and was accomplished under Cooperative Agreement number W911NF-18-2-0048. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the US Government. The US Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Author contributions All authors contributed substantially to this work. X.Y., C.P. and B.Z. wrote the manuscript with contributions from all authors. M.S., C.P. and B.Z. supervised the project.

Competing interests The authors declare no competing interests.

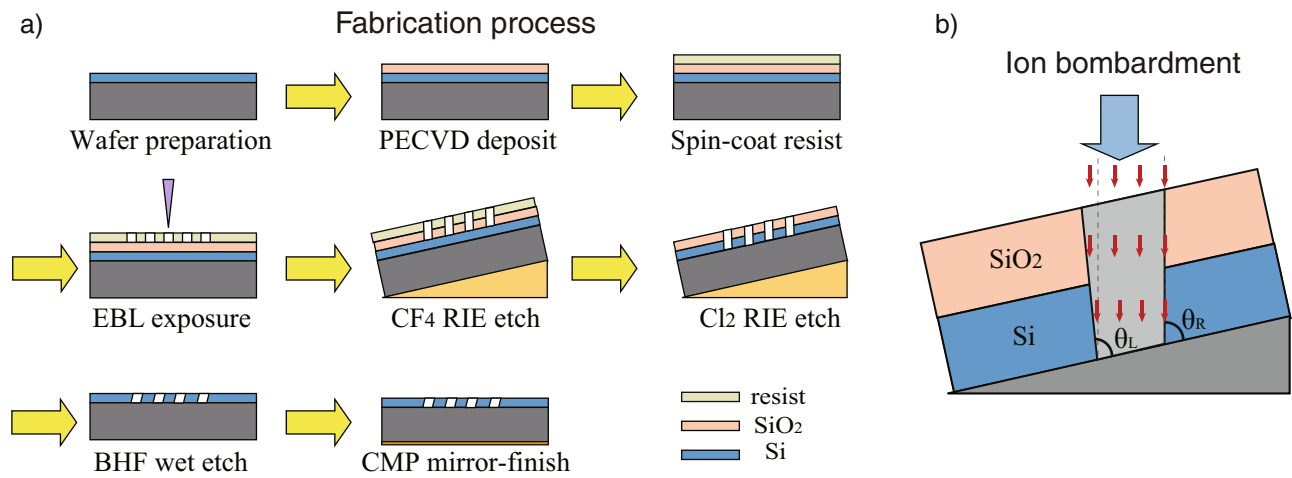
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2181-4>.

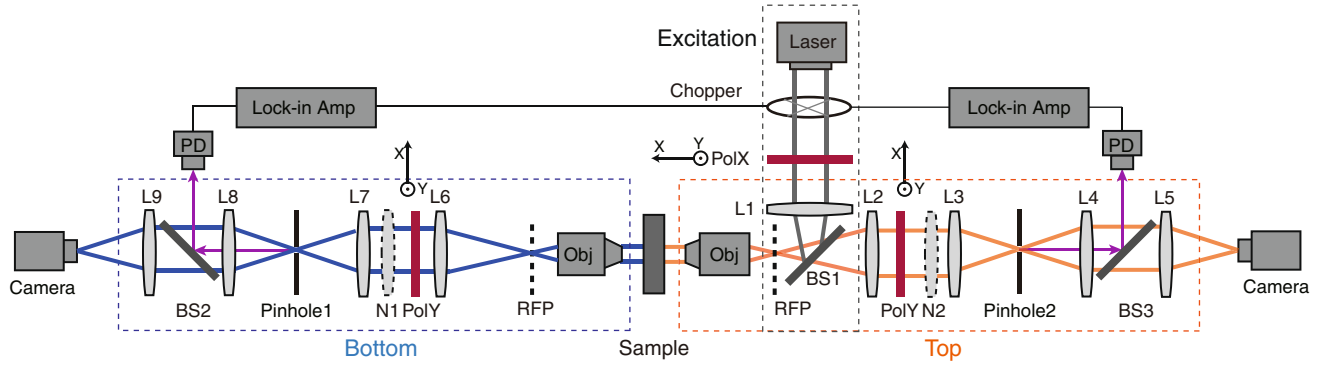
Correspondence and requests for materials should be addressed to C.P.

Peer review information *Nature* thanks Yuri Kivshar, Mikael Rechtsman and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

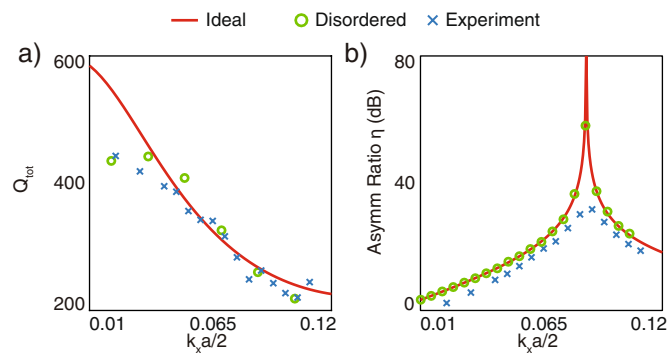


Extended Data Fig. 1 | Sample fabrication. **a**, Step-by-step flow chart of the fabrication process. **b**, Schematics of the customized RIE process. EBL, electron-beam lithography; PECVD, plasma-enhanced chemical vapour deposition; CMP, chemical-mechanical polishing.

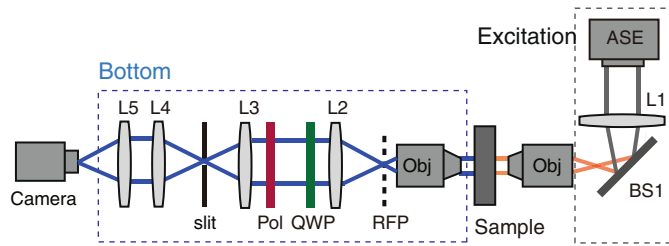


Extended Data Fig. 2 | Experimental setup used to measure the asymmetry ratio η . The setup is capable of both near- and far-field measurements. The focal lengths of lenses L2, L3, L4 and L5 are 150 mm, 100 mm, 75 mm and 75 mm,

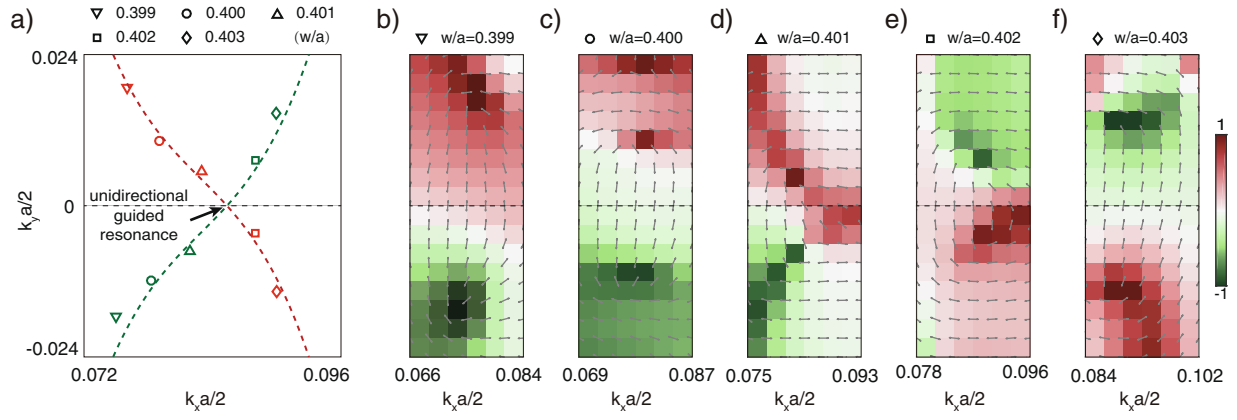
respectively. RFP, rear focal plane; PD, photodetector; Obj, objective; Pol, polarizer; Amp, amplifier; BS, beam splitter. N1 and N2 denote the movable lenses used to achieve near-field imaging.



Extended Data Fig. 3 | Experimental and simulation results for disordered samples. a, Experimentally extracted Q_{tot} (blue) compared with simulation results for samples with (green) and without (red) disorder. **b,** Measured asymmetry ratio η (blue) compared with simulation results for samples with (green) and without (red) disorder.

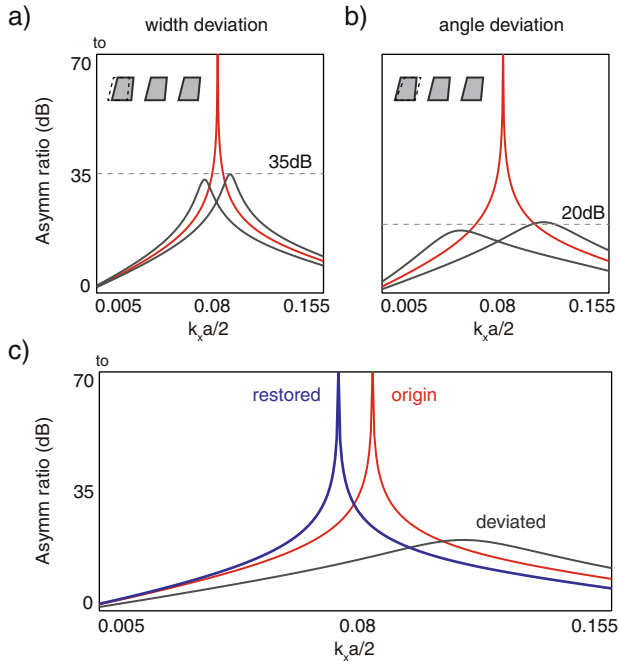


Extended Data Fig. 4 | Experimental setup used for polarimetry measurements. An amplified spontaneous emission (ASE) source excites the resonances in the sample. Scattered light is recorded by a camera under six different combinations of a polarizer (Pol) and a QWP. The focal lengths of lenses L2, L3, L4 and L5 are 150 mm, 100 mm, 75 mm and 75 mm, respectively.

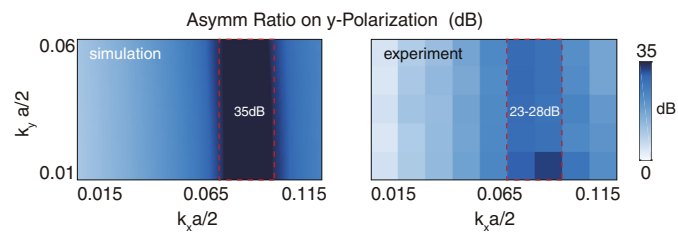


Extended Data Fig. 5 | Experimental observation of the evolution of half-integer charges. **a**, UGR as the merging point between two half-integer charges. **b–f**, Measured ellipticity ρ of the resonances in five samples with

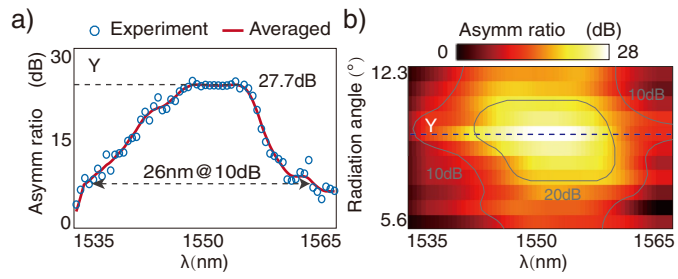
slightly different air-gap widths w , ranging from $w/a = 0.399$ (**b**) to 0.403 (**f**). Dark red ($\rho = 1$) and dark green ($\rho = -1$) colours indicate the locations of the LCP and RCP resonances, which are also half-integer topological charges.



Extended Data Fig. 6 | Robustness of UGRs against parameter variations. **a**, Device performance when the air-gap widths deviate by ± 2.5 nm from the perfect design. **b**, Device performance when the etching angle deviates by $\pm 1^\circ$ from the perfect design (grey). **c**, The UGR is restored if the etching angle deviates by -1° from the perfect design and the air-gap width changes to $w = 365$ nm.



Extended Data Fig. 7 | Asymmetry ratio for modes near UGRs. Simulated (left) and measured (right) asymmetry ratios η for resonances close to the UGR in momentum space.



Extended Data Fig. 8 | Prospects of using UGRs as grating couplers.

a. Asymmetry ratio η between upward and downward radiation intensities for a fixed out-coupling angle of 9° . The maximum reaches 27.7 dB near the UGR and remains high (above 10 dB) over a bandwidth of 26 nm. **b.** Highly directional emission is observed over a wide range of excitation wavelengths and for different out-coupling angles. The fibre-to-waveguide loss is not measured.

Extended Data Table 1 | Comparison of different mechanisms used to achieve highly directional radiation

Mechanism	Asymmetry ratio (dB)		Maximum coupling efficiency (%)		Ref
	numerical	experimental ¹	numerical	experimental ^{1,2}	
non-resonant blazing effect	8.7	×	80.1	×	[38]
non-resonant blazing effect	20	×	up to 99	×	[39]
non-resonant blazing effect	20	> 7.96	up to 99	86.2	[40]
non-resonant blazing effect	20	×	up to 99	×	[41]
dual-layer guided resonance	20	> 10.6	95	92	[36]
dual-layer guided resonance	20	×	95	×	[42]
dual-layer guided resonance	21	×	99.2	×	[43]
dual-layer guided resonance	8.9	×	70	×	[44]
UGR	70	27.7	×	×	this work

Data from refs. ³⁸⁻⁴⁴ and from this work.
¹Lower bound on the measurement value.
²Not including taper loss.

Strongly correlated electrons and hybrid excitons in a moiré heterostructure

<https://doi.org/10.1038/s41586-020-2191-2>

Received: 1 November 2019

Accepted: 27 February 2020

Published online: 13 April 2020

 Check for updates

Yuya Shimazaki^{1,3}✉, Ido Schwartz^{1,3}, Kenji Watanabe², Takashi Taniguchi², Martin Kroner¹ & Ataç Imamoğlu^{1,3}

Two-dimensional materials and their heterostructures constitute a promising platform to study correlated electronic states, as well as the many-body physics of excitons. Transport measurements on twisted graphene bilayers have revealed a plethora of intertwined electronic phases, including Mott insulators, strange metals and superconductors^{1–5}. However, signatures of such strong electronic correlations in optical spectroscopy have hitherto remained unexplored. Here we present experiments showing how excitons that are dynamically screened by itinerant electrons to form exciton-polarons^{6,7} can be used as a spectroscopic tool to investigate interaction-induced incompressible states of electrons. We study a molybdenum diselenide/hexagonal boron nitride/molybdenum diselenide heterostructure that exhibits a long-period moiré superlattice, as evidenced by coherent hole-tunnelling-mediated avoided crossings of an intralayer exciton with three interlayer exciton resonances separated by about five millielectronvolts. For electron densities corresponding to half-filling of the lowest moiré subband, we observe strong layer pseudospin paramagnetism, demonstrated by an abrupt transfer of all the (roughly 1,500) electrons from one molybdenum diselenide layer to the other on application of a small perpendicular electric field. Remarkably, the electronic state at half-filling of each molybdenum diselenide layer is resilient towards charge redistribution by the applied electric field, demonstrating an incompressible Mott-like state of electrons. Our experiments demonstrate that optical spectroscopy provides a powerful tool for investigating strongly correlated electron physics in the bulk and paves the way for investigating Bose–Fermi mixtures of degenerate electrons and dipolar excitons.

Van der Waals heterostructures incorporating transition metal dichalcogenide (TMD) bilayers open up new avenues for exploring strong correlations using transport and optical spectroscopy. In contrast to similar structures in III–V semiconductors, these heterostructures exhibit possibilities for exotic material combinations, the creation of moiré superlattices exhibiting narrow electronic bands^{8–11} and strong binding of spatially separated interlayer excitons¹². Recently, transport experiments in twisted bilayer graphene demonstrated strongly correlated electron physics in a single system^{1–5}, ranging from superconductivity to a Mott insulator state as the filling factor ν is varied. In fact, this system realizes a two-dimensional (2D) Fermi–Hubbard model on a triangular lattice with a tunable electron density.

In parallel, optical spectroscopy in van der Waals heterostructures have revealed the prevalence of many-body hybrid light–matter states, termed exciton-polarons^{6,7}, in the excitation spectra of electron- or hole-doped monolayers. Advances in material quality and device fabrication has led to the observation of moiré physics of non-interacting excitons in TMD heterobilayers^{13–16}. The potential of this new system for many-body physics was recently revealed in a demonstration of a long-lived interlayer exciton condensate¹⁷. Here we describe

experiments in a heterostructure incorporating a molybdenum diselenide/hexagonal boron nitride/molybdenum diselenide (MoSe₂/hBN/MoSe₂) homobilayer that in several ways combine the principal developments in these two fields to demonstrate interaction-induced incompressible states of electrons. We provide an unequivocal demonstration of the hybridization of inter- and intralayer excitons by coherent hole tunnelling^{11,15,18} between the two MoSe₂ layers: the avoided crossings in the optical reflection not only show the formation of dipolar excitons with strong optical coupling but also reveal the existence of moiré bands of interlayer excitons. We then demonstrate that intralayer exciton-polaron resonances provide a sensitive tool to investigate correlated electronic states in the bulk. Equipped with this spectroscopic tool, we observe strong layer pseudospin paramagnetism^{19,20} and an incompressible Mott-like state of electrons at half-filling of each layer.

Device structure and basic characterization

Figure 1a shows a schematic of the device structure. By using a double-gate structure, we control the electric field E_z and the chemical potential μ of the device independently (see Methods for details).

¹Institute for Quantum Electronics, ETH Zürich, Zurich, Switzerland. ²National Institute for Materials Science, Tsukuba, Japan. ³These authors contributed equally: Yuya Shimazaki, Ido Schwartz.

✉e-mail: yuyas@phys.ethz.ch; imamoglu@phys.ethz.ch

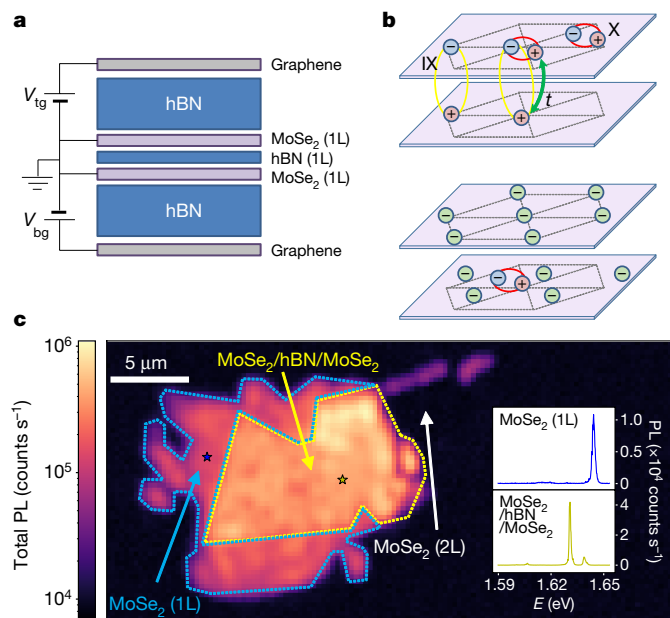


Fig. 1 | Device structure and basic characterization. **a**, Schematic of the device structure. V_{tg} (V_{bg}) is the applied voltage to the top (bottom) gate. **b**, Schematic of coupled inter- and intralayer exciton (top) and electrons in a moiré lattice probed by the intralayer exciton (bottom). Purple planes correspond to MoSe₂ layers and dashed lines indicate a moiré lattice. The pink (light blue) circles with a + (–) sign indicate holes (electrons) forming excitons, and the electron–hole pair enclosed by the red (yellow) ellipse indicates intralayer (interlayer) exciton. The green double arrow in the top panel indicates tunnel coupling (t) of holes through a monolayer hBN barrier. The light green circles in the bottom panel indicate electrons filling the moiré lattice. **c**, Spatial map of integrated photoluminescence from 1.59 eV to 1.65 eV. The blue and yellow dashed lines indicate the boundary of the monolayer MoSe₂ and MoSe₂/hBN/MoSe₂ regions, respectively. The inset shows representative PL spectra of monolayer MoSe₂ and MoSe₂/hBN/MoSe₂ measured at the positions indicated with the blue and yellow stars in the main figure, respectively. 1L, monolayer; 2L, bilayer.

Figure 1b (top) is a schematic image of a dipolar exciton formed by coherent coupling of an interlayer exciton (IX) and an intralayer exciton (X) via hole tunnelling. Figure 1b (bottom) shows a sketch of electrons in a moiré lattice, probed by intralayer excitons.

Figure 1c shows a spatial map of total photoluminescence (PL) from the device. Here both the top and bottom gate voltages are at zero volts. We observe PL from regions with monolayer MoSe₂, but not from bilayer MoSe₂, where two MoSe₂ flakes are in direct contact (the region around the point indicated by the white arrow in Fig. 1c). In contrast, the MoSe₂/hBN/MoSe₂ region shows bright PL. This indicates that the heterostructure becomes a direct-bandgap system owing to the reduction of the interlayer hybridization of the valence bands at the Γ point²¹, due to the presence of monolayer hBN. Typical PL spectra of the monolayer MoSe₂ and the MoSe₂/hBN/MoSe₂ region are shown in the inset of Fig. 1c: there are pronounced intralayer exciton luminescence peaks in both regions. We observe two distinct exciton peaks in the MoSe₂/hBN/MoSe₂ region. This strain-induced energy difference between the PL from the top and bottom layers varies across the sample (Supplementary Information section 2).

Coherent interlayer hole tunnelling and dipolar excitons

We first analyse the electric field (E_z) dependence of the elementary optical excitations of the MoSe₂/hBN/MoSe₂ region in the absence of itinerant electrons or holes. To this end, we scan the top and bottom

gate voltages together to change E_z while keeping the system in the charge-neutral regime. The obtained PL spectrum is depicted in Fig. 2a: using the top (V_{tg}) and bottom (V_{bg}) gate voltage dependence, we determine that the PL spectra around 1.632 eV and 1.640 eV stem from an intralayer exciton in the top and bottom layer (X_{top} and X_{bot}), respectively. For high values of $|E_z|$ depicted in the top and bottom parts of the colour-coded PL spectrum, we observe PL lines with a strong E_z dependence: we identify these PL lines as originating from interlayer excitons with a large dipole moment leading to a sizeable Stark shift.

The spectra for a positive (negative) V_{tg} regime correspond to the interlayer exciton IX_{\uparrow} (IX_{\downarrow}), which has a hole in the bottom (top) layer and an electron in the top (bottom) layer. The associated dipole moment of the interlayer exciton changes its polarity for $V_{tg} \approx 0$. By extrapolating the IX_{\uparrow} and IX_{\downarrow} PL lines and finding their crossing point, we estimate the energy difference between the inter- and the intralayer exciton resonances at $E_z = 0$, which allows us to determine their binding energy difference to be 0.1 ± 0.01 eV (Supplementary Information section 6).

Figure 2b shows the differential reflectance ($\Delta R/R_0$) spectrum obtained for the same range of gate voltage scan as that of Fig. 2a. Here $\Delta R/R_0 \equiv (R - R_0)/R_0$, with R and R_0 denoting the reflectance signal from the MoSe₂/hBN/MoSe₂ region and background reflectance, respectively. In accordance with the PL data (Fig. 2a), we find X_{top} and X_{bot} resonances around 1.632 eV and 1.640 eV, respectively. Moreover, for $V_{tg} \gtrsim 7.5$ V ($V_{tg} \lesssim -7.5$ V), we observe IX_{\uparrow} (IX_{\downarrow}) hybridizing exclusively with X_{top} (X_{bot}). Figure 2c, d shows magnified plots of the regions highlighted with blue and green dashed lines in Fig. 2b, confirming avoided crossing of an intralayer exciton with multiple interlayer excitons. We first note that the observation of a sizeable reflection signal from IX_{\uparrow} away from the avoided crossing suggests that it is possible to resonantly excite long-lived interlayer excitons in these structures. The hybridization of IX_{\uparrow} lines with X_{top} , together with the lack of an avoided crossing with X_{bot} in Fig. 2c, unequivocally shows that avoided crossings originate exclusively from the coherent hole tunnelling schematically shown in Fig. 2e. Our observation, proving that the hole tunnel coupling is much larger than that of the electron, is consistent with the band alignment expected from first-principles band-structure calculations²². This conclusion is also confirmed by the data depicted in Fig. 2d, where avoided crossing originates from the coherent-hole-tunnelling-induced hybridization of IX_{\downarrow} and X_{bot} schematically shown in Fig. 2f.

One of the most remarkable features of the spectra depicted in Fig. 2c, d is the existence of multiple avoided crossings associated with three interlayer exciton resonances separated in energy by about 5 meV. This interlayer exciton fine structure demonstrates the existence of a moiré superlattice^{8,10,11,13–16}, originating from a small twist angle between the two MoSe₂ layers (Methods). The presence of an hBN tunnel barrier strongly suppresses the strength of the associated moiré potential, rendering it sizeable for only the interlayer excitons⁸.

Charge configuration detection by exciton-polaron spectroscopy

The presence of itinerant charges drastically alters the optical excitation spectrum²³. Recent theoretical and experimental work established that the modified spectrum originates from dynamical screening of excitons by electrons or holes^{6,7}, leading to the formation of a lower-energy attractive polaron (AP) branch. Concurrently, the exciton resonance evolves into a repulsive polaron (RP) (Supplementary Information section 1). The strong sensitivity of the RP resonance energy to changes in electron density renders it an ideal spectroscopic tool for sensing the electron density n in the same layer^{24,25}. The strain-induced resonance energy difference between X_{top} and X_{bot} , ensuring different energies for the corresponding RP_{top} and RP_{bot} , together with the much weaker sensitivity of RP_{top} (RP_{bot}) to the electron density n_b (n_t) in the bottom (top) layer, allows us to determine the charging configuration of the two layers simultaneously. As we are predominantly interested in the

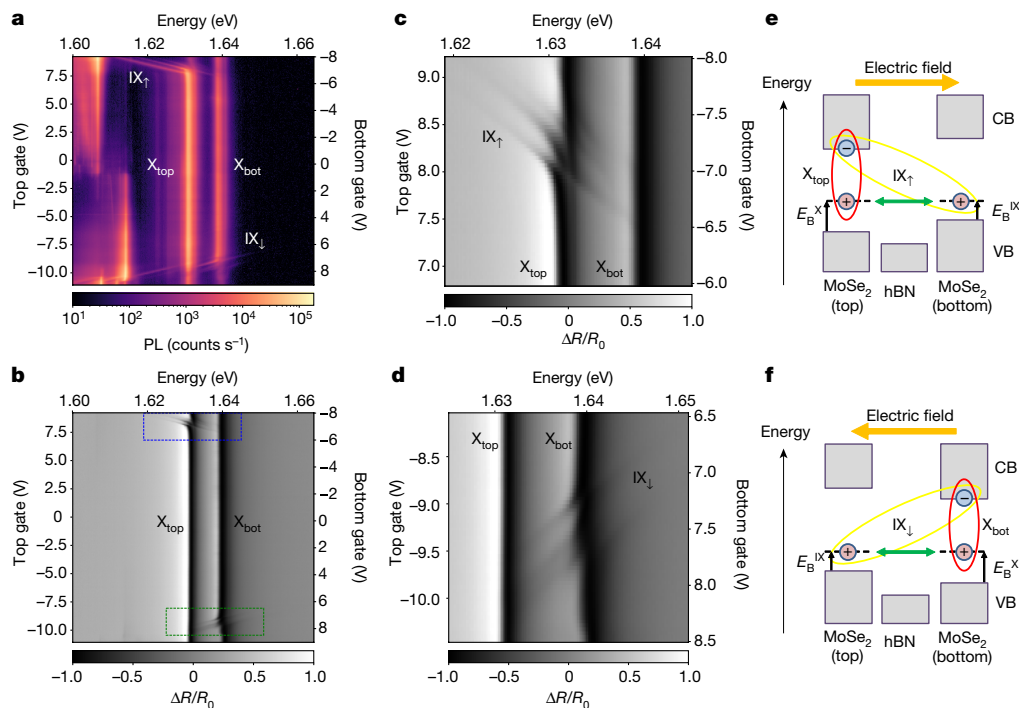


Fig. 2 | Electric field dependence of PL and differential reflectance at charge neutrality. **a, b**, Gate dependence of PL (**a**) and differential reflectance (**b**) of MoSe₂/hBN/MoSe₂. Top and bottom gate voltages are scanned together to tune the electric field at a constant chemical potential (scanned along the dashed line L4 shown in Fig. 3c). The intensity of the PL in **a** is shown on a log scale. **c, d**, Magnified plots of **b**. The corresponding region of **c** and **d** is indicated by the blue and green dashed rectangles in **b**, respectively.

e, f, Schematic of the energy bands and the exciton energy alignment under electric fields where the excitons are in resonance. CB and VB denote the conduction and valence bands, respectively, and E_B^{IX} denotes the exciton binding energy. Green double arrows represent hole tunnelling. Panels **e** and **f** are the cases for opposite direction of electric field, and correspond to **c** and **d**, respectively.

low-carrier-density regime where the quasiparticle (bare exciton) weight of the RP resonance is close to unity, we will refer to it as the exciton resonance.

Figure 3a, b shows the gate voltage dependence of $\Delta R/R_0$ at $E = 1.632$ eV and $E = 1.640$ eV, which correspond to the top (X_{top}) and bottom (X_{bot}) intralayer exciton resonance energy in the charge-neutral regime, respectively. The insets to these figures show a line cut through the dispersive neutral exciton reflection spectrum, indicating the exciton energies at which we monitor $\Delta R/R_0$. As a small increase of electron density from about 0 to 1×10^{11} cm⁻² results in a change of $\Delta R/R_0$ from about -1 to ≥ 0 , the blue regions in Fig. 3a, b correspond to the charge-neutral regime of each layer. The red and white regions in turn, correspond to the electron- or hole-doped regime of each layer. This all-optical determination of the charge map of the bilayer provides an invaluable tool for monitoring the bulk properties of 2D materials.

To enhance the sensitivity of the charge map and to visualize the charge configuration of both layers at the same time, we evaluate and overlay the derivative of $\Delta R/R_0$ ($d(\Delta R/R_0)/dE$) obtained for both layers (see also Supplementary Information section 7). The resulting charge map, depicted in Fig. 3c, is closely reminiscent of the charging diagram used to characterize gate-defined quantum dots²⁶. The blue regions in Fig. 3c correspond to gate voltages where the charge configuration changes, allowing us to clearly separate the regions (t, b) where the top or bottom layer is neutral (t = i or b = i), electron doped (t = n or b = n) or hole doped (t = p or b = p).

We show the typical gate voltage dependence of $\Delta R/R_0$ in Fig. 3d, e obtained while scanning the two gates along the lines L1 and L2, indicated in Fig. 3a, b, respectively. In both plots, we confirm the emergence of the AP resonance and the associated blueshift of the exciton or RP energy around the charge configuration transition points, confirming the assignment obtained from $d(\Delta R/R_0)/dE$ in Fig. 3c.

Figure 3f shows the gate voltage scan along L3, indicated in Fig. 3c, where we fixed V_{bg} at 4 V and scanned V_{tg} . As we increase the chemical potential by sweeping V_{tg} from negative to positive, we find that initially the bottom layer gets electron doped, because the (single particle) conduction band energy of the bottom layer is lower than that of the top layer for the chosen V_{bg} . For $V_{tg} \approx 4$ V, electrons are introduced into the top layer as well. In the absence of interactions, we would expect the electron density in the bottom layer to become independent of a further increase of V_{tg} , due to screening of V_{tg} by the free electrons in the top layer. We observe different behaviour in Fig. 3f: increasing V_{tg} results in a decrease and eventually total depletion of electrons from the bottom layer. The underlying physics is understood by considering the intralayer exchange interaction, which is maximal if all electrons occupy a single layer. If the ensuing reduction in total repulsive Coulomb energy exceeds the kinetic energy cost of having all electrons in a single layer, layer polarization is favoured, leading to the observed depletion of electrons from the bottom layer. This phenomenon, termed negative compressibility, was previously observed in transport experiments in bilayer semiconductor systems²⁷.

Interaction-induced incompressible states

The results we present in the ‘Coherent interlayer hole tunnelling and dipolar excitons’ section establish the existence of a moiré superlattice for interlayer excitons. The absence of coherent electron tunnelling, however, indicates that the corresponding electronic moiré subbands in the top and bottom layers do not hybridize. Therefore, our homobilayer realizes a rather unique system exhibiting flat bands with layer and valley-spin degrees of freedom. Moreover, negative compressibility (‘Charge configuration detection by exciton-polaron spectroscopy’ section) indicates that electron–electron interactions are prominent

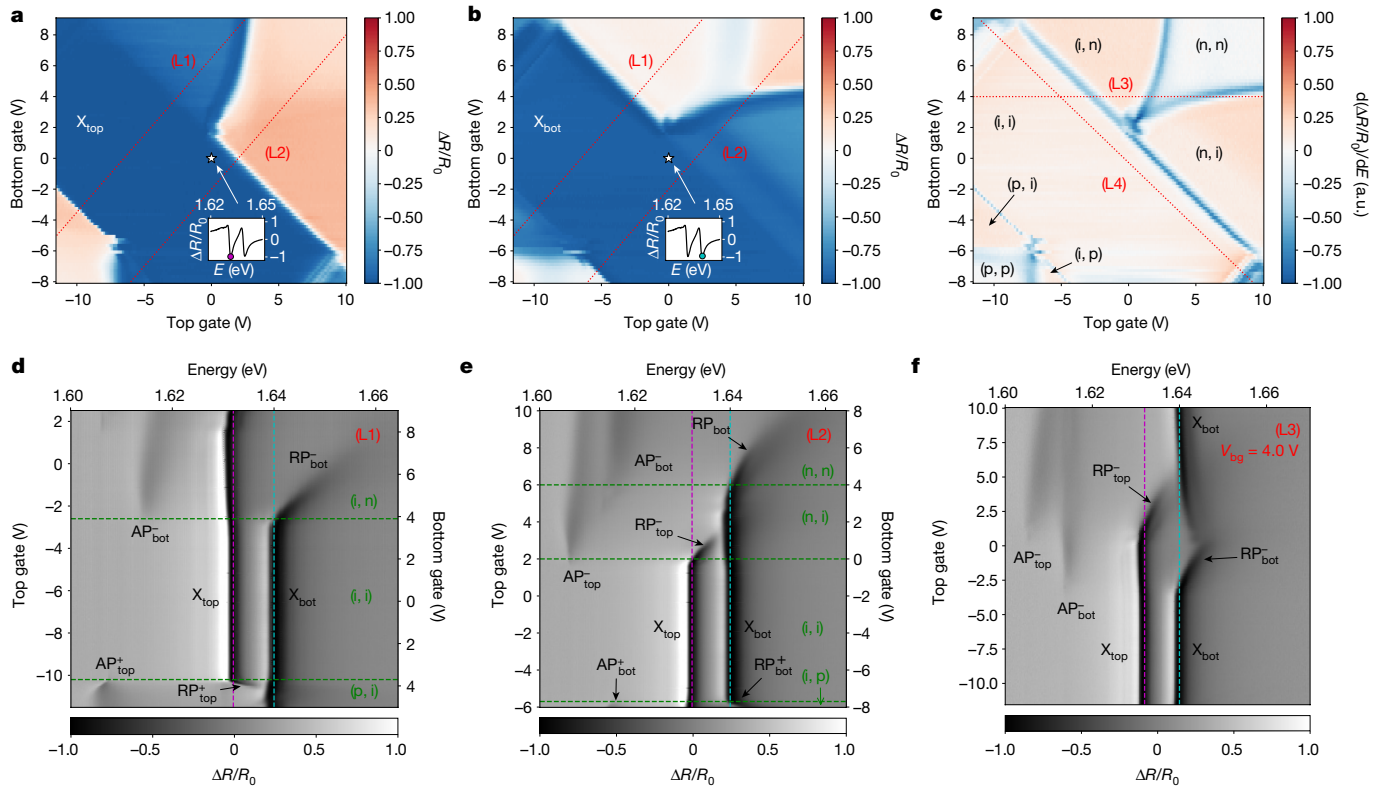


Fig. 3 | Gate dependence of differential reflectance spectrum of MoSe₂/hBN/MoSe₂. **a, b**, Two gates dependence maps of differential reflectance around the top (**a**) and bottom (**b**) intralayer exciton resonances ($E = 1.632$ eV and $E = 1.640$ eV, respectively). The insets of **a** and **b** show the differential reflectance spectrum at $(V_{\text{tg}}, V_{\text{bg}}) = (0 \text{ V}, 0 \text{ V})$ (indicated with the white stars in the maps). The magenta and cyan dots in the insets indicate the points where $E = 1.632$ eV and $E = 1.640$ eV, respectively. **c**, Charge configuration diagram obtained by derivative of the differential reflectance spectrum with respect to energy (sum of the derivatives at $E = 1.632$ eV and $E = 1.640$ eV). The charge configuration for each layer is indicated by p, i and n, which correspond to hole

doped (p), neutral (i) and electron doped (n), and shown in the order of (top, bottom). **d–f**, Gate dependence of differential reflectance along the dashed lines L1 (**d**), L2 (**e**) and L3 (**f**) shown in **a–c**. Magenta and cyan dashed lines indicate the top ($E = 1.632$ eV) and bottom ($E = 1.640$ eV) exciton resonance energies, respectively. AP_L^{C} and RP_L^{C} stand for intralayer attractive and repulsive polarons, where L is ‘top’ or ‘bot’ stands for top or bottom layer, and C is + or – stands for hole or electron as Fermi sea carriers. Charge configuration is written in green together with the green dashed lines indicating the charge configuration transition point.

even at relatively high electron densities ($n \approx 1 \times 10^{12} \text{ cm}^{-2}$) where several moiré bands in one layer are occupied. In this section, we explore electron correlation effects by focusing on the low- n regime of the charging map (Fig. 3c) where the (i, i), (i, n), (n, i) and (n, n) regions coalesce. The high sensitivity of the exciton/RP resonance energy, as well as the AP oscillator strength, to changes in electron density once again forms the backbone of our investigation.

Figure 4a shows the gate voltage dependence of $\Delta R/R_0$ at energy of 1.6320 eV. The choice of the energy at which we monitor $\Delta R/R_0$ maximizes the sensitivity to the shift of X_{top} and is indicated by the magenta point in the inset of Fig. 4a. We now choose the horizontal and vertical voltage axes to be $V_E = 0.5V_{\text{tg}} - 0.5V_{\text{bg}}$ and $V_{\mu} = (7/15)V_{\text{tg}} + (8/15)V_{\text{bg}}$. With this choice, vertical (V_{μ} axis) and horizontal (V_E axis) cuts through the reflectance map leave E_z and μ unchanged, respectively. Remarkably, Fig. 4a shows a periodic modulation of the reflectance as a function of V_{μ} , particularly in the low- n regime. Moreover, the modulation of the X_{top} and X_{bot} reflectance are correlated and symmetric with respect to the $V_E = -1 \text{ V}$ axis (Supplementary Fig. 4), indicating that for this value of V_E , the conduction band edge energy of the two layers is aligned.

To gain further insight, we first determine the repulsive polaron resonance energy ($E_{X_{\text{top}}}$) by fitting the reflectance spectrum with a dispersive Lorentzian line shape (Supplementary Information section 3) and then plot the derivative of $E_{X_{\text{top}}}$ with respect to V_{μ} in Fig. 4b. The resulting map shows a remarkable checkerboard pattern that is complementary for the top and bottom layers (Supplementary Fig. 4).

As the blueshift of X_{top} (X_{bot}) resonance while increasing V_{μ} corresponds to filling of electrons in the top (bottom) layer, the complementary checkerboard patterns in Fig. 4b and Supplementary Fig. 4f indicate a layer-by-layer filling of electrons²⁸.

The observed periodicity in Fig. 4 indicates the existence of moiré subbands for electrons. In anticipation of the subsequent discussion, we define a layer filling factor ν_L (where L is ‘top’ or ‘bot’, indicating top or bottom layer, respectively) such that $\nu_L = 1/2$ corresponds to 1 electron per moiré unit cell of a single layer, and a total filling factor ν as $\nu = \nu_{\text{top}} + \nu_{\text{bot}}$. From a capacitive model of our device, we determine that $\nu = 1/2$ coincides with electron density of $n = 2 \times 10^{11} \text{ cm}^{-2}$. At this low electron density, the r_s parameter, which describes the ratio of interaction energy to kinetic energy, is estimated to be $r_s \geq 14$. The density periodicity corresponds to a moiré superlattice constant of $\lambda_{\text{moiré}} = 24 \text{ nm}$ by assuming a triangular superlattice (see Methods for the estimation of r_s and $\lambda_{\text{moiré}}$). We indicate the values of ν corresponding to $\nu = 1/2, 1, 3/2, 2$ with blue dashed lines in Fig. 4b. We confirmed that the same periodicity also appears in the AP photoluminescence intensity (Supplementary Fig. 4g, h).

Figure 5a, b shows the V_E dependence of $\Delta R/R_0$ for fixed V_{μ} for $\nu = 1/2$ and $\nu = 1$, respectively. In Fig. 5a, we find an abrupt shift of exciton energy together with complete oscillator strength transfer between AP_{top} and AP_{bot} , demonstrating that the electrons are completely transferred from one layer to the other. Figure 5c, e shows the extracted X_{bot} and X_{top} energies around $\nu = 1/2$. Remarkably, the abrupt jump in excitonic

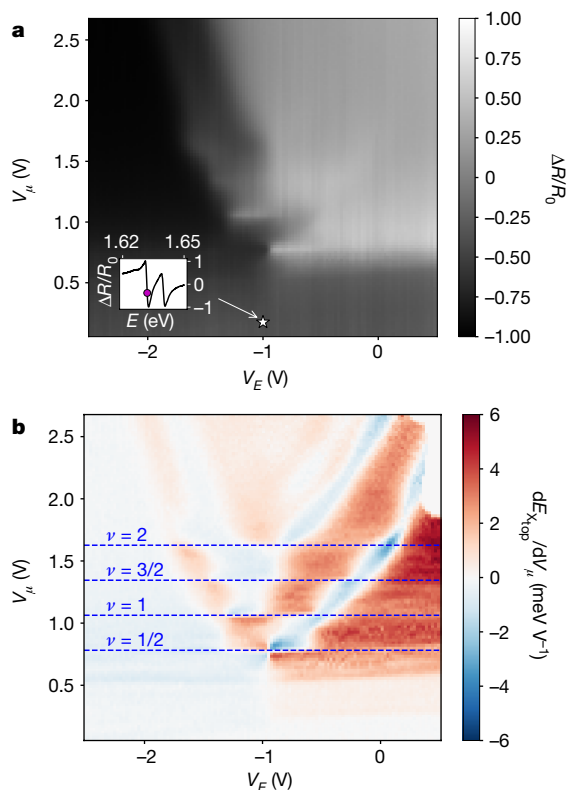


Fig. 4 | Gate dependence of intralayer exciton resonance in the low-electron-density regime. **a**, Gate dependence map of differential reflectance around the top intralayer exciton resonance ($E = 1.6320$ eV). The inset of **a** shows the differential reflectance spectrum at $(V_E, \nu_\mu) = (-1\text{ V}, 0.175\text{ V})$ (indicated with the white star in the map). **b**, Gate dependence map of the top intralayer exciton resonance energy differentiated by ν_μ .

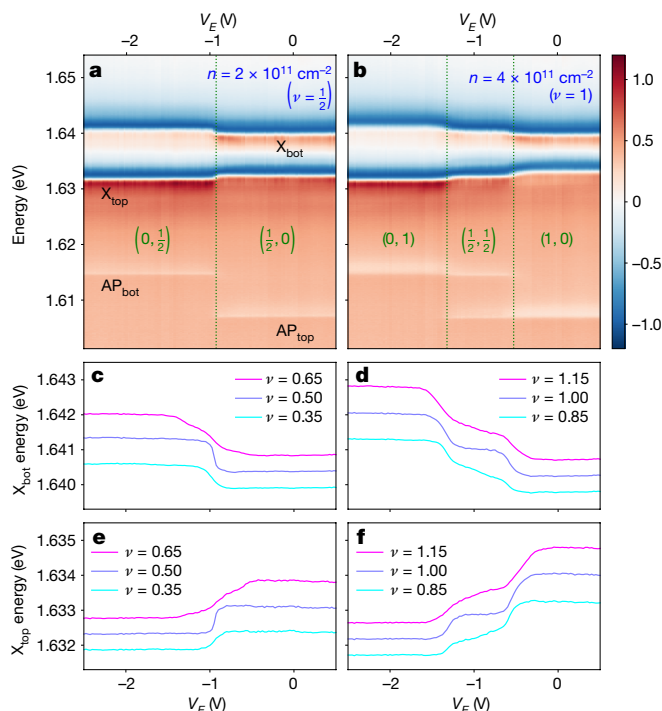


Fig. 5 | Electric field dependence of differential reflectance spectrum in the low-electron-density regime. **a, b**, Electric field (V_E) dependence of differential reflectance spectrum for fixed ν_μ at $\nu = 1/2$ and $\nu = 1$. The charge configuration of the top and bottom layers is indicated by $(\nu_{\text{top}}, \nu_{\text{bot}})$ in green. **c, d**, V_E dependence of X_{bot} resonance energy around total filling of $\nu = 1/2$ (**c**) and $\nu = 1$ (**d**). **e, f**, V_E dependence of X_{top} resonance energy around total filling of $\nu = 1/2$ (**e**) and $\nu = 1$ (**f**). In **c–f**, the cyan curves are without offset and other curves are displaced by 0.5 meV (purple) and 1.0 meV (magenta). **g, h**, Schematic of the charge configuration with density of states at a filling of $\nu = 1/2$ (**g**) and $\nu = 1$ (**h**).

resonance is pronounced at $\nu = 1/2$, and smeared out for both lower ($\nu = 0.35$) and higher filling factors ($\nu = 0.65$). These measurements show that abrupt transfer of practically all of the roughly 1,500 electrons within the region we monitor optically is linked to the emergence of an interaction-induced incompressible state in the lowest moiré sub-band at $\nu = 1/2$ filling. As the filling factor is increased or decreased towards $\nu = 1/2$, the electronic system shows an ever-stronger layer pseudospin paramagnetism, due to the enhanced role of interactions, but otherwise exhibits a continuous interlayer transfer of electrons as a function of E_z that would be expected from a compressible state. Close to $\nu = 1/2$, there is a phase transition to an incompressible state that can be accommodated in either the top or the bottom layer (Fig. 5g). Remarkably, interlayer charge transfer takes place upon changing V_E by only 1.9 mV: the corresponding change in the single-particle energy detuning between the top and bottom layers ($26\text{ }\mu\text{eV}$) is much smaller than $k_B T$ ($360\text{ }\mu\text{eV}$; where k_B is the Boltzmann constant and T is the temperature) (Supplementary Information section 5).

Figure 5b shows that for $\nu = 1$, $\Delta R/R_0$ is characterized by three plateau-like regions. We attribute the abrupt jumps in the excitonic resonance energy to the transition from $(\nu_{\text{top}}, \nu_{\text{bot}}) = (0, 1)$, through $(1/2, 1/2)$, to $(1, 0)$ configurations (Fig. 5h). This explanation is confirmed by the corresponding changes in the oscillator strength of the AP resonances of the top and bottom layers. In the $(1, 0)$ and $(0, 1)$ configurations, we measure a reflectance signal either from AP_{top} or AP_{bot} , consistent with full-layer polarization. In the $(1/2, 1/2)$ configuration, we find the oscillator strength of AP_{top} and AP_{bot} to be identical and equal to half the value obtained under $(1, 0)$ for AP_{top} . The extracted excitonic resonance energy for X_{bot} and X_{top} around $\nu = 1$ is shown in Fig. 5d, f, respectively. The plateau structure of the $(1/2, 1/2)$ state with abrupt changes of electron density difference between the layers at $V_E = -1.3\text{ V}$ and $V_E = -0.5\text{ V}$ is clearly visible at $\nu = 1$, but is smeared out for both lower ($\nu = 0.85$) and higher fillings ($\nu = 1.15$). The emergence of the stabilized $(1/2, 1/2)$ plateau at $\nu = 1$ strongly suggests that there is mutual stabilization of the incompressible electronic state due to the interlayer electron–electron interactions. The energy gap of the incompressible $(1/2, 1/2)$ state is estimated to be 5.5 meV (Supplementary Information section 6). The reflectance data for higher fillings ($\nu = 3/2$ and $\nu = 2$) are

shown in the Supplementary Information (Supplementary Fig. 5): in stark contrast to the (1/2, 1/2) configuration at $\nu = 1$, a plateau at the (1, 1) electron configuration is missing at $\nu = 2$ filling, indicating that the state with the corresponding integer fillings is not sufficiently stabilized by the interlayer interactions.

Finally, we also perform the measurement under a perpendicular magnetic field $B_z = 7$ T (Supplementary Information). In Supplementary Fig. 6, we find that the plateau structure observed for $\nu = 2$ under full valley polarization of electrons is identical to that observed for $B_z = 0$ T, although the total number of electronic states per moiré subband is halved due to giant valley-spinsusceptibility of electrons in MoSe_2 (ref. ²⁴). This observation shows that the incompressibility is determined by filling of each moiré site by a single electron, irrespective of its (valley) degeneracy. The observation supports that our identification of $n = 2 \times 10^{11} \text{ cm}^{-2}$, yielding half-filling of a single-layer moiré subband. We also find that the reflectance spectrum does not show any valley polarization at $B_z = 7$ T for the $\nu \leq 1/2$ state. However, the resilience against valley polarization is not sufficient to claim antiferromagnetic order.

Discussion

The experiments we describe in the ‘Interaction-induced incompressible states’ section demonstrate the existence of Mott-like incompressible electronic states for half-filling of the lowest moiré subband. Unlike previous reports^{13–16}, our experiments are carried out for a long moiré superlattice lattice constant of $\lambda_{\text{moiré}} = 24$ nm and an r_s parameter of $r_s \geq 14$. The weakness of the moiré potential stemming from the hBN layer separating the two MoSe_2 layers in turn ensures that the on-site interaction strength is larger than the depth of the moiré potential. In this sense, the homobilayer system realizes a rather unique regime that goes beyond the standard Fermi–Hubbard model.

In addition to establishing twisted TMD homobilayers as a promising system for investigating Mott–Wigner physics^{10,29–31} originating from strong electronic correlations, our experiments open up new avenues for exploring interactions between dipolar excitons and electrons confined to flat bands. In particular, the structure we analysed could be used to realize and study Bose–Fermi mixtures consisting of degenerate electrons strongly interacting with an exciton condensate generated by resonant laser excitation. The phase diagram of such a mixture is not fully understood³², but is expected to provide a rich playground for many-body physics³³.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2191-2>

1. Cao, Y. et al. Unconventional superconductivity in magic-angle graphene superlattices. *Nature* **556**, 43–50 (2018).
2. Yankowitz, M. et al. Tuning superconductivity in twisted bilayer graphene. *Science* **363**, 1059–1064 (2019).

3. Liu, X. et al. Spin-polarized correlated insulator and superconductor in twisted double bilayer graphene. Preprint at <http://arxiv.org/abs/1903.08130> (2019).
4. Sharpe, A. L. et al. Emergent ferromagnetism near three-quarters filling in twisted bilayer graphene. *Science* **365**, 605–608 (2019).
5. Lu, X. et al. Superconductors, orbital magnets and correlated states in magic-angle bilayer graphene. *Nature* **574**, 653–657 (2019).
6. Sidler, M. et al. Fermi polaron-polaritons in charge-tunable atomically thin semiconductors. *Nat. Phys.* **13**, 255–261 (2017).
7. Efimkin, D. K. & MacDonald, A. H. Many-body theory of trion absorption features in two-dimensional semiconductors. *Phys. Rev. B* **95**, 035417 (2017).
8. Yu, H., Liu, G.-B., Tang, J., Xu, X. & Yao, W. Moiré excitons: from programmable quantum emitter arrays to spin-orbit-coupled artificial lattices. *Sci. Adv.* **3**, e1701696 (2017).
9. Wu, F., Lovorn, T. & MacDonald, A. Topological exciton bands in moiré heterojunctions. *Phys. Rev. Lett.* **118**, 147401 (2017).
10. Wu, F., Lovorn, T., Tutuc, E. & MacDonald, A. H. Hubbard model physics in transition metal dichalcogenide moiré bands. *Phys. Rev. Lett.* **121**, 026402 (2018).
11. Ruiz-Tijerina, D. A. & Fal’ko, V. I. Interlayer hybridization and moiré superlattice minibands for electrons and excitons in heterobilayers of transition-metal dichalcogenides. *Phys. Rev. B* **99**, 125424 (2019).
12. Rivera, P. et al. Interlayer valley excitons in heterobilayers of transition metal dichalcogenides. *Nat. Nanotechnol.* **13**, 1004–1015 (2018).
13. Seyler, K. L. et al. Signatures of moiré-trapped valley excitons in $\text{MoSe}_2/\text{WS}_2$ heterobilayers. *Nature* **567**, 66–70 (2019).
14. Tran, K. et al. Evidence for moiré excitons in van der Waals heterostructures. *Nature* **567**, 71–75 (2019).
15. Alexeev, E. M. et al. Resonantly hybridized excitons in moiré superlattices in van der Waals heterostructures. *Nature* **567**, 81–86 (2019); correction **572**, E8 (2019).
16. Jin, C. et al. Observation of moiré excitons in WSe_2/WS_2 heterostructure superlattices. *Nature* **567**, 76–80 (2019); correction **569**, E7 (2019).
17. Wang, Z. et al. Evidence of high-temperature exciton condensation in two-dimensional atomic double layers. *Nature* **574**, 76–80 (2019).
18. Gerber, I. C. et al. Interlayer excitons in bilayer MoS_2 with strong oscillator strength up to room temperature. *Phys. Rev. B* **99**, 035443 (2019).
19. Zheng, L., Ortalan, M. W. & Das Sarma, S. Exchange instabilities in semiconductor double-quantum-well systems. *Phys. Rev. B* **55**, 4506–4515 (1997).
20. Ezawa, Z. F. *Quantum Hall Effects: Field Theoretical Approach and Related Topics* (World Scientific, 2000).
21. Zhang, Y. et al. Direct observation of the transition from indirect to direct bandgap in atomically thin epitaxial MoSe_2 . *Nat. Nanotechnol.* **9**, 111–115 (2014).
22. Özçelik, V. O., Azadani, J. G., Yang, C., Koester, S. J. & Low, T. Band alignment of two-dimensional semiconductors for designing heterostructures with momentum space matching. *Phys. Rev. B* **94**, 035125 (2016).
23. Xu, X., Yao, W., Xiao, D. & Heinz, T. F. Spin and pseudospins in layered transition metal dichalcogenides. *Nat. Phys.* **10**, 343–350 (2014).
24. Back, P. et al. Giant paramagnetism-induced valley polarization of electrons in charge-tunable monolayer MoSe_2 . *Phys. Rev. Lett.* **118**, 237404 (2017).
25. Smoleński, T. et al. Interaction-induced Shubnikov–de Haas oscillations in optical conductivity of monolayer MoSe_2 . *Phys. Rev. Lett.* **123**, 097403 (2019).
26. Hanson, R., Kouwenhoven, L. P., Petta, J. R., Tarucha, S. & Vandersypen, L. M. K. Spins in few-electron quantum dots. *Rev. Mod. Phys.* **79**, 1217–1265 (2007).
27. Eisenstein, J. P., Pfeiffer, L. N. & West, K. W. Compressibility of the two-dimensional electron gas: measurements of the zero-field exchange energy and fractional quantum Hall gap. *Phys. Rev. B* **50**, 1760–1778 (1994).
28. Hunt, B. M. et al. Direct measurement of discrete valley and orbital quantum numbers in bilayer graphene. *Nat. Commun.* **8**, 948 (2017).
29. Imada, M., Fujimori, A. & Tokura, Y. Metal–insulator transitions. *Rev. Mod. Phys.* **70**, 1039–1263 (1998).
30. Camjayi, A., Haule, K., Dobrosavljević, V. & Kotliar, G. Coulomb correlations and the Wigner–Mott transition. *Nat. Phys.* **4**, 932–935 (2008).
31. Zarenia, M., Neilson, D. & Peeters, F. M. Inhomogeneous phases in coupled electron-hole bilayer graphene sheets: charge density waves and coupled Wigner crystals. *Sci. Rep.* **7**, 11510 (2017).
32. Ludwig, D., Floerchinger, S., Moroz, S. & Wetterich, C. Quantum phase transition in Bose–Fermi mixtures. *Phys. Rev. A* **84**, 033629 (2011).
33. Laussy, F. P., Kavokin, A. V. & Shelykh, I. A. Exciton-polariton mediated superconductivity. *Phys. Rev. Lett.* **104**, 106402 (2010).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Device fabrication

The device structure is shown in Extended Data Fig. 1. All MoSe₂, graphene and hBN flakes were obtained by mechanical exfoliation of bulk crystals. The flakes were assembled together using the dry transfer technique³⁴ in an argon-filled glove box. The crystal axis of top and bottom MoSe₂ layers were aligned to be close to 0° using a tear-and-stack technique³⁵. We expect to have the effect of moiré potential from a small twist angle of MoSe₂ layers, but not between MoSe₂ and hBN layers due to the large lattice constant mismatch ($a_{\text{MoSe}_2} = 3.32 \text{ \AA}$ (ref. 22) and $a_{\text{hBN}} = 2.51 \text{ \AA}$ (ref. 36)). The metal electrodes to graphene layers are formed by Ti/Au (5 nm/145 nm). The contact to the bottom graphene gate is formed by Cr/Au (3 nm/147 nm) using the 1D contact technique³⁴ by etching the hBN layer with reactive ion etching using CHF₃/O₂ as mixture gas.

Optical spectroscopy

The PL measurements were performed using a HeNe laser (633 nm) as an excitation source. The reflectance measurements were performed using a single-mode fibre-coupled broadband light-emitting diode with a centre wavelength of 760 nm and a bandwidth of 20 nm. In both PL and reflectance measurements, we used a long working distance apochromatic objective lens with a numerical aperture of 0.65 (attocube LT-APO/LWD/VISIR/0.65). All optical spectroscopy measurements were performed at cryogenic temperature ($T \approx 4 \text{ K}$). All data in the main text were obtained at zero magnetic field, $B_z = 0 \text{ T}$.

Capacitance parameters

The thickness of the top and bottom hBN layers was estimated to be 63.3 nm and 52.7 nm, respectively, from the white light reflectance spectrum. We used 3.7 as the dielectric constant for hBN³⁷. We expect the uncertainty to be about 10% for both hBN thickness and the dielectric constant, which gives around 14% error in the calculation of carrier density.

Estimation of r_s parameter and moiré periodicity from electron density

The r_s parameter (Wigner–Seitz radius divided by the effective Bohr radius) that describes the ratio between kinetic energy to Coulomb energy is given by $r_s = 1/a_B^* \sqrt{\pi n}$ (ref. 19). Here $a_B^* = 4\pi\epsilon\epsilon_0\hbar^2/m_e^*e^2 = 0.91 \text{ nm}$ is the effective Bohr radius in an encapsulated MoSe₂ monolayer, m_e^* is the effective electron mass, $\epsilon = (\epsilon_{\parallel, \text{MoSe}_2} + \epsilon_{\parallel, \text{hBN}})/2$ is the lattice dielectric constant¹⁹, ϵ_0 is the vacuum permittivity, \hbar is the reduced Planck constant and e is the elementary charge. For the calculation we used $m_e^* = 0.7m_e$ (ref. 38), where m_e is the bare electron mass and the in-plane dielectric constants $\epsilon_{\parallel, \text{MoSe}_2} = 17.1$ and $\epsilon_{\parallel, \text{hBN}} = 6.93$ (ref. 37). For an electron density of $n = 2 \times 10^{11} \text{ cm}^{-2}$ ($\nu = 1/2$), we obtain $r_s \approx 14$, which is reduced to $r_s \approx 7$ at density of $n = 8 \times 10^{11} \text{ cm}^{-2}$ ($\nu = 2$). We emphasize that the above values for the r_s parameter are underestimated. At these low densities, the interelectron separation, $r_0 = 1/\sqrt{\pi n} > 10 \text{ nm}$, is much larger than the MoSe₂ layer thickness $d_{\text{MoSe}_2} \approx 0.7 \text{ nm}$. In this limit³⁹, the screening of the interactions is dominated by the hBN dielectric; therefore, a better approximation is to take the dielectric constant to be $\epsilon = \epsilon_{\parallel, \text{hBN}}$, which results in $r_s \approx 24$ for an electron density of $n = 2 \times 10^{11} \text{ cm}^{-2}$ ($\nu = 1/2$) and $r_s \approx 12$ for $n = 8 \times 10^{11} \text{ cm}^{-2}$ ($\nu = 2$).

The moiré superlattice lattice constant $\lambda_{\text{moiré}}$ is estimated from the following relation by assuming a triangular moiré superlattice: $A_{\text{moiré}} = (\sqrt{3}/2)\lambda_{\text{moiré}}^2 = 1/n_{\text{moiré}}$, where $A_{\text{moiré}}$ is the moiré superlattice unit cell area and $n_{\text{moiré}}$ is the electron density that corresponds to one electron occupation per moiré superlattice unit cell. We used $n_{\text{moiré}} = 2 \times 10^{11} \text{ cm}^{-2}$, which is obtained from the periodicity of the differential reflectance modulation as we discussed in the main text, which gives $\lambda_{\text{moiré}} \approx 24 \text{ nm}$. The twist angle θ between MoSe₂ layers is then estimated to be $\theta = 0.8^\circ$ using the following relation: $\lambda_{\text{moiré}} = a_{\text{MoSe}_2}/\theta$, where $a_{\text{MoSe}_2} = 3.32 \text{ \AA}$ is the lattice constant of MoSe₂ (ref. 22).

Verification of moiré periodicity estimation using interlayer exciton fine-structure analysis

Here we estimate the moiré periodicity from the fine-structure energy scale of the interlayer exciton. In the limit when the moiré potential is weak compared with the kinetic energy scale, umklapp scattering by the moiré potential is the dominant origin of the fine structure of the interlayer exciton. In this limit, the energy difference of the two lowest energy interlayer excitons is approximately described by the kinetic energy of the interlayer exciton at the momentum $q_{\text{moiré}} = 4\pi/(\sqrt{3}\lambda_{\text{moiré}})$, where $\lambda_{\text{moiré}}$ is moiré periodicity^{9,11}. Using the energy splitting of the first- and second-lowest energy interlayer exciton $\Delta E_{1,2}$, $\lambda_{\text{moiré}}$ is obtained using

$$\Delta E_{1,2} \approx \frac{\hbar^2 q_{\text{moiré}}^2}{2M} \quad (1)$$

$$\lambda_{\text{moiré}} \approx \frac{2h}{\sqrt{6M\Delta E_{1,2}}} \quad (2)$$

where the total mass of exciton is $M = m_e^* + m_h^* = 1.3m_e$, m_e^* is the effective electron mass, m_h^* is the effective hole mass ($m_e^* = 0.7m_e$ from ref. 38 and $m_h^* = 0.6m_e$ from ref. 21) and h is the Planck constant. From Fig. 2c in the main text, the energy splitting is $\Delta E_{1,2} \approx 4.4 \text{ meV}$, which corresponds to $\lambda_{\text{moiré}} \approx 19 \text{ nm}$. At a different spot, from Supplementary Fig. 3d, the energy splitting is $\Delta E_{1,2} \approx 3.7 \text{ meV}$, which corresponds to $\lambda_{\text{moiré}} \approx 20 \text{ nm}$. Both values are qualitatively similar to what we obtained from the density periodicity of the reflectance signal ($\lambda_{\text{moiré}} \approx 24 \text{ nm}$ from Fig. 4b and $\lambda_{\text{moiré}} \approx 25 \text{ nm}$ from Supplementary Fig. 10). We note that the third-lowest energy interlayer exciton line observed in reflectance is not captured well by this simple model. In reality, the energy spacing of the interlayer excitons depends on the magnitude of moiré potential, and the analysis we present here could only yield a rather rough estimate of the moiré periodicity.

Effect of strain on moiré periodicity

A moiré superlattice emerges from a twist angle, lattice constant difference or a combination of the two. In the main text, we attributed our superlattice to be originating from this twist angle. Here we discuss how much strain-induced lattice constant modification changes the moiré periodicity. Moiré periodicity for twist angle of θ and lattice constant difference ratio of two layers δ (ref. 40), which corresponds to strain difference of two layers, is expressed as:

$$\lambda_{\text{moiré}} = \frac{a_{\text{MoSe}_2}(1 + \delta)}{\sqrt{2(1 + \delta)(1 - \cos\theta) + \delta^2}} \quad (3)$$

Extended Data Fig. 2a plots the relation between the biaxial strain difference (δ) and twist angle (θ) for fixed $\lambda_{\text{moiré}}$ using equation (3) ($a_{\text{MoSe}_2} = 3.32 \text{ \AA}$). In the main text, we have shown that there is an intralayer exciton energy difference between the top and bottom layers, which is about 8 meV. By assuming this energy difference to arise from the strain difference between the two layers, we extract the amount of strain difference to be on the order of about 0.1–0.25%. We base our rough estimate on values reported in literature: uniaxial strain MoS₂: 70 meV %⁻¹ (ref. 41), 45 meV %⁻¹ (ref. 42), 48 meV %⁻¹ (ref. 43); biaxial strain MoSe₂: 33 meV %⁻¹ (ref. 44). For $\lambda_{\text{moiré}} \approx 24 \text{ nm}$, the moiré periodicity within this strain range is dominated by twist angle and hardly modified by strain (only 1.4% reduction of $\lambda_{\text{moiré}}$ for 0.25% strain assuming fixed twist angle of 0.8°, as we show in Extended Data Fig. 2b). However, to introduce a moiré periodicity of about 24 nm with only strain difference, more than 1.3% of strain difference is required, which is unlikely from the magnitude of the energy splitting of the intralayer excitons. The spectrum at another spot, which we show in Supplementary Information section 2, exhibits 5 meV energy splitting of the top- and

bottom-layer intralayer excitons. From this energy difference, the amount of strain difference is estimated to be in the order of about 0.07–0.15%, which results in only 0.5% of $\lambda_{\text{moiré}}$ reduction, assuming fixed twist angle of 0.8°.

First-principles calculation of the tunnel coupling and the moiré potential

We perform density functional theory (DFT) calculations using Quantum ESPRESSO⁴⁵. We use projector-augmented-wave pseudopotentials with a generalized gradient approximation (Perdew–Burke–Ernzerhof functional) from PSLibrary 1.0.0 (ref. ⁴⁶). To reduce computation cost, we neglect spin–orbit interaction and used non-relativistic pseudopotentials. Computations are performed with 48 cores on a high-performance computing cluster.

We take lattice structure parameters for MoSe₂ from ref. ⁴⁷, which gives a lattice constant $a_{\text{MoSe}_2} = 3.32 \text{ \AA}$. The lattice constant of hBN given in ref. ³⁶ is $a_{\text{hBN}} = 2.504 \text{ \AA}$. We uniformly stretched the hBN lattice by –0.56% and set the lattice constant of hBN as $a_{\text{hBN}}^{\text{stretch}} = 2.49 \text{ \AA}$ to form a 3×3 MoSe₂ and 4×4 hBN commensurate supercell. We take the interlayer distance of MoSe₂ and hBN from ref. ⁴⁸, and set the thickness of vacuum layer between MoSe₂ and hBN as $d_{\text{Se-hBN}} = 3.36 \text{ \AA}$. We use a plane-wave cutoff energy of 60 Rydberg and a charge density cutoff energy of 480 Rydberg. The Brillouin zone is sampled with a $9 \times 9 \times 1$ k-point grid. We use a slab geometry with a 30-Å-thick vacuum layer.

Extended Data Fig. 3a–c shows the actual R-stacked (0° twist angle) MoSe₂/hBN/MoSe₂ supercell for different lattice displacement configurations that we use for the calculation. $\alpha = \text{h, X, M}$ denote hexagon centre, chalcogen and metal site of TMDs, and R_h^α denotes the specific lattice displacement for R-stacking where the α site of top MoSe₂ is aligned with the h site of the bottom MoSe₂. Extended Data Fig. 3d–f shows the calculated band structure shown with the mini-Brillouin zone. We take the vacuum level as the energy reference (0 eV). Owing to zone folding, the K and K' points of MoSe₂ come to the γ point of the mini-Brillouin zone (the K point of hBN comes to the κ point of the mini-Brillouin zone). Therefore, the lowest conduction bands and the highest valence bands at the γ point are the ones from MoSe₂. As the interlayer hybridization effect is substantially smaller than the directly contacting TMD heterostructure^{8–10}, these band structures look almost the same except for a slightly visible energy splitting of R_h^h displacement. Though the energy modulation is substantially smaller, the calculation shows about 0.5 meV modulation for the lowest conduction band and 5 meV modulation for the highest valence band dependent on the lattice displacement (Extended Data Table 1). In addition, the calculation shows 11-meV tunnel splitting of the valence band edge at R_h^h displacement, which is qualitatively in good agreement with experimentally observed avoided crossings of intralayer and interlayer excitons mediated by hole tunnelling. We emphasize that these results are rather qualitative, and in reality, it depends on many factors such as interlayer distance, outer dielectric environment and so on.

Data availability

The data that support the findings of this study are available in the ETH Research Collection (<http://hdl.handle.net/20.500.11850/399579>).

34. Wang, L. et al. One-dimensional electrical contact to a two-dimensional material. *Science* **342**, 614–617 (2013).
35. Kim, K. et al. van der Waals heterostructures with high accuracy rotational alignment. *Nano Lett.* **16**, 1989–1995 (2016).
36. Catellani, A., Posternak, M., Baldereschi, A. & Freeman, A. J. Bulk and surface electronic structure of hexagonal boron nitride. *Phys. Rev. B* **36**, 6105–6111 (1987).
37. Laturia, A., Van de Put, M. L. & Vandenberghe, W. G. Dielectric properties of hexagonal boron nitride and transition metal dichalcogenides: from monolayer to bulk. *npj 2D Mater. Appl.* **2**, 6 (2018).
38. Larentis, S. et al. Large effective mass and interaction-enhanced Zeeman splitting of K-valley electrons in MoSe₂. *Phys. Rev. B* **97**, 201407 (2018).
39. Rytova, N. S. The screened potential of a point charge in a thin film. *Moscow Univ. Phys. Bull.* **22**, 18–21 (1967).
40. Lu, C.-P., Li, G., Watanabe, K., Taniguchi, T. & Andrei, E. Y. MoS₂: choice substrate for accessing and tuning the electronic properties of graphene. *Phys. Rev. Lett.* **113**, 156804 (2014).
41. He, K., Poole, C., Mak, K. F. & Shan, J. Experimental demonstration of continuous electronic structure tuning via strain in atomically thin MoS₂. *Nano Lett.* **13**, 2931–2936 (2013).
42. Conley, H. J. et al. Bandgap engineering of strained monolayer and bilayer MoS₂. *Nano Lett.* **13**, 3626–3630 (2013).
43. Zhu, C. R. et al. Strain tuning of optical emission energy and polarization in monolayer and bilayer MoS₂. *Phys. Rev. B* **88**, 121301 (2013).
44. Frisenda, R. et al. Biaxial strain tuning of the optical properties of single-layer transition metal dichalcogenides. *npj 2D Mater. Appl.* **1**, 10 (2017).
45. Giannozzi, P. et al. QUANTUM ESPRESSO: a modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **21**, 395502 (2009).
46. Dal Corso, A. Pseudopotentials periodic table: from H to Pu. *Comput. Mater. Sci.* **95**, 337–350 (2014).
47. Rasmussen, F. A. & Thygesen, K. S. Computational 2D materials database: electronic structure of transition-metal dichalcogenides and oxides. *J. Phys. Chem. C* **119**, 13169–13183 (2015).
48. Zollner, K., Faria, P. E. Jr & Fabian, J. Proximity exchange effects in MoSe₂ and WSe₂ heterostructures with CrI₃: twist angle, layer, and gate dependence. *Phys. Rev. B* **100**, 085128 (2019).

Acknowledgements We acknowledge discussions with E. Demler, R. Schmidt, T. Smolenski, A. Popert and P. Knüppel. This work was supported by the Swiss National Science Foundation (SNSF) under grant number 200021-178909/1 and the European Research Council (ERC) Advanced Investigator Grant (POLTDES). Y.S. acknowledges support from the Japan Society for the Promotion of Science (JSPS) overseas research fellowships. K.W. and T.T. acknowledge support from the Elemental Strategy Initiative conducted by MEXT, Japan, A3 Foresight by JSPS and CREST (grant number JPMJCR15F3) and JST.

Author contributions Y.S. and I.S. carried out the measurements. Y.S. designed and fabricated the sample. M.K. helped to prepare the experimental setup. K.W. and T.T. grew the hBN crystal. Y.S. performed DFT calculation. Y.S., I.S. and A.I. wrote the manuscript. A.I. supervised the project.

Competing interests The authors declare no competing interests.

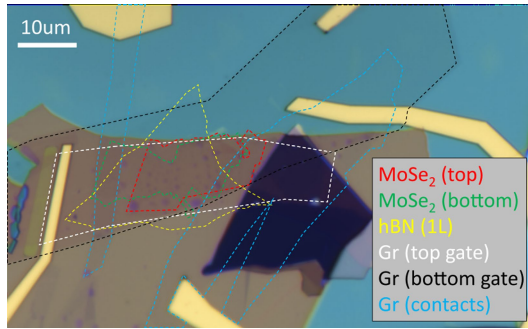
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2191-2>.

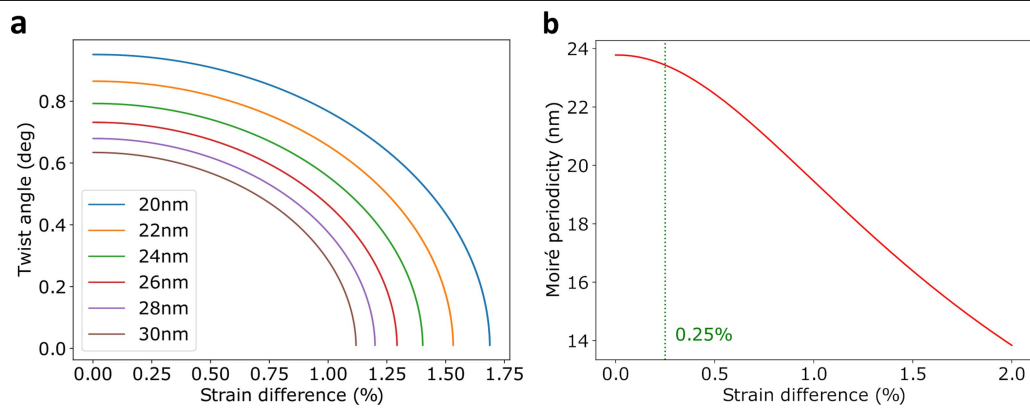
Correspondence and requests for materials should be addressed to Y.S. or A.I.

Peer review information Nature thanks Wang Yao and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

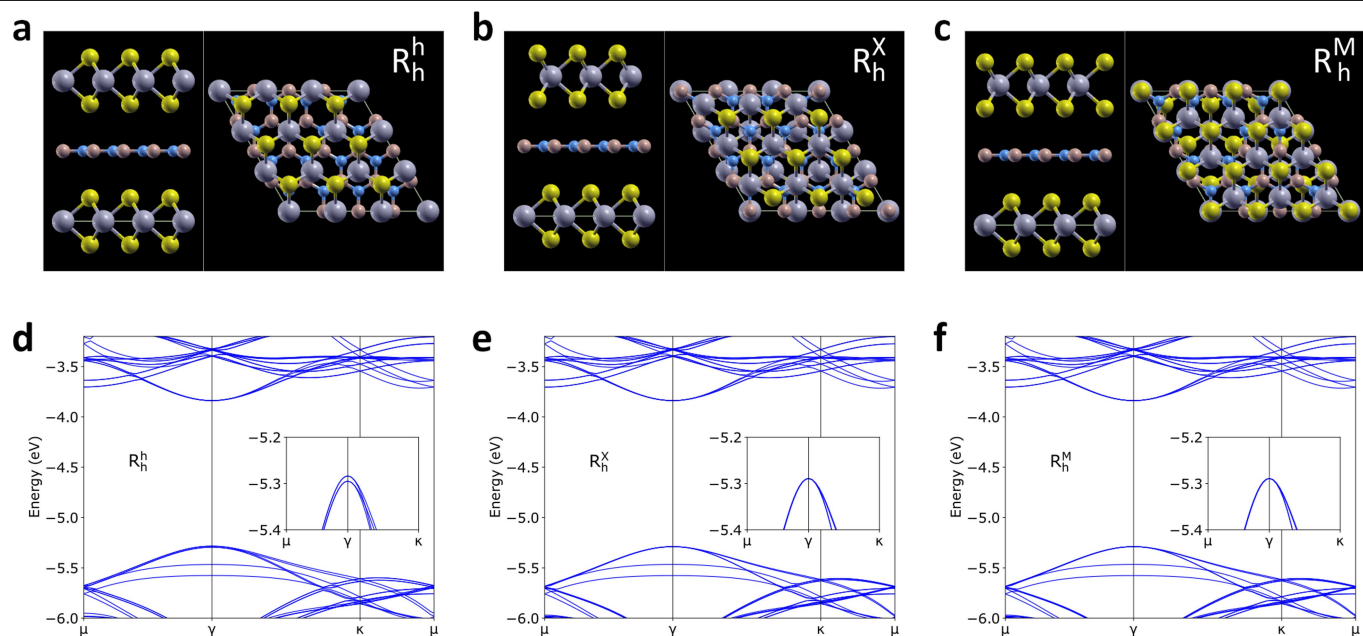
Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | Optical microscope image of the device. The border of each flake is highlighted with dashed lines, and the material is indicated in the grey box with the corresponding colour. Gr, graphene.



Extended Data Fig. 2 | Effect of twist angle and strain on moiré periodicity. a, Plot of the relation between twist angle and strain difference which gives same moiré periodicity ($\lambda_{\text{moiré}}$), shown for $\lambda_{\text{moiré}}$ from 20 to 30 nm. **b,** Strain difference dependence of moiré periodicity for a fixed twist angle of 0.8°.



Extended Data Fig. 3 | Band structure of R-stacked MoSe₂/hBN/MoSe₂ heterostructure obtained from DFT calculation. a–c, The side and top view of the supercell for R_h^h (a), R_h^x (b) and R_h^M (c) used for the calculation. **d–f,** The

calculated band structure of R-stacked MoSe₂/hBN/MoSe₂ for R_h^h (d), R_h^x (e) and R_h^M (f) lattice displacement. The insets show the magnified plot of the valence bands around the γ point.

Extended Data Table 1 | DFT calculation results for MoSe₂/hBN/MoSe₂

Lattice displacement	R_h^h	R_h^x	R_h^M
Lowest conduction band edge (eV)	-3.8389	-3.8383	-3.8383
Highest valence band edge (eV)	-5.2835	-5.2889	-5.2890

Enhanced ferroelectricity in ultrathin films grown directly on silicon

<https://doi.org/10.1038/s41586-020-2208-x>

Received: 15 July 2019

Accepted: 27 January 2020

Published online: 22 April 2020

 Check for updates

Suraj S. Cheema^{1,13}✉, Daewoong Kwon^{2,12,13}, Nirmaan Shanker^{1,2}, Roberto dos Reis³, Shang-Lin Hsu^{3,7}, Jun Xiao⁴, Haigang Zhang⁵, Ryan Wagner⁵, Adhiraj Datar^{1,2}, Margaret R. McCarter⁶, Claudy R. Serrao², Ajay K. Yadav², Golnaz Karbasian², Cheng-Hsiang Hsu², Ava J. Tan², Li-Chen Wang¹, Vishal Thakare¹, Xiang Zhang⁴, Apurva Mehta⁸, Evguenia Karapetrova⁹, Rajesh V Chopdekar¹⁰, Padraic Shafer¹⁰, Elke Arenholz^{10,11}, Chenming Hu², Roger Proksch⁵, Ramamoorthy Ramesh^{1,6}, Jim Ciston³ & Sayeef Salahuddin^{2,7}✉

Ultrathin ferroelectric materials could potentially enable low-power perovskite ferroelectric tetragonality logic and nonvolatile memories^{1,2}. As ferroelectric materials are made thinner, however, the ferroelectricity is usually suppressed. Size effects in ferroelectrics have been thoroughly investigated in perovskite oxides—the archetypal ferroelectric system³. Perovskites, however, have so far proved unsuitable for thickness scaling and integration with modern semiconductor processes⁴. Here we report ferroelectricity in ultrathin doped hafnium oxide (HfO₂), a fluorite-structure oxide grown by atomic layer deposition on silicon. We demonstrate the persistence of inversion symmetry breaking and spontaneous, switchable polarization down to a thickness of one nanometre. Our results indicate not only the absence of a ferroelectric critical thickness but also enhanced polar distortions as film thickness is reduced, unlike in perovskite ferroelectrics. This approach to enhancing ferroelectricity in ultrathin layers could provide a route towards polarization-driven memories and ferroelectric-based advanced transistors. This work shifts the search for the fundamental limits of ferroelectricity to simpler transition-metal oxide systems—that is, from perovskite-derived complex oxides to fluorite-structure binary oxides—in which ‘reverse’ size effects counterintuitively stabilize polar symmetry in the ultrathin regime.

Ferroelectric materials exhibit stable states of collectively ordered electrical dipoles whose polarization can be reversed under an applied electric field⁵. Consequently, ultrathin ferroelectrics are of great technological interest for high-density electronics, particularly field-effect transistors and non-volatile memories². However, ferroelectricity is typically suppressed at the scale of a few nanometres in the ubiquitous perovskite oxides⁶. First-principles calculations predict six unit cells to be the critical thickness in perovskite ferroelectrics¹ owing to incomplete screening of depolarization fields³. Atomic-scale ferroelectricity in perovskites often fails to demonstrate polarization switching^{7,8}, which is crucial for applications. Furthermore, attempts to synthesize ferroelectric perovskite films on silicon^{9,10} are impeded by chemical incompatibility^{4,11} and the high temperatures required for epitaxial growth. Since the discovery of ferroelectricity in HfO₂-based thin films in 2011¹², fluorite-structure binary oxides (fluorites) have attracted considerable interest¹³ because they enable low-temperature synthesis and conformal growth in three-dimensional structures on silicon^{14,15},

thereby overcoming many of the issues that restrict its perovskite counterparts in terms of complementary metal-oxide-semiconductor (CMOS) compatibility and thickness scaling¹⁶. Considering the extensive implications for future computing^{2,17,18}, achieving ferroelectricity in sub-2-nm-thick doped-HfO₂ is highly desirable for realizing ultra-scaled CMOS-compatible ferroelectric-based devices beyond the 5 nm technology node¹⁹.

Here we demonstrate ferroelectricity in ultrathin (1 nm thick) Hf_{0.8}Zr_{0.2}O₂ (HZO), grown by low-temperature atomic layer deposition (ALD) on silicon. Second harmonic generation and advanced scanning probe techniques establish the presence of inversion symmetry breaking and switchable electric polarization, respectively. Not only is ferroelectricity stabilized in ultrathin HZO, but spectroscopic and diffraction signatures of its fluorite-structure symmetry also indicate enhanced polar distortion in the ultrathin regime. Such size effects in this fluorite-structure system do not occur in its perovskite counterparts⁶, which can be understood from symmetry considerations. In

¹Department of Materials Science and Engineering, University of California, Berkeley, CA, USA. ²Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA. ³National Center for Electron Microscopy, Molecular Foundry, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ⁴Nanoscale Science and Engineering Center, University of California, Berkeley, CA, USA. ⁵Asylum Research, Oxford Instruments, Santa Barbara, CA, USA. ⁶Department of Physics, University of California, Berkeley, Berkeley, CA, USA. ⁷Materials Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ⁸Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA, USA. ⁹Advanced Photon Source, Argonne National Laboratory, Argonne, IL, USA. ¹⁰Advanced Light Source, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. ¹¹Cornell High Energy Synchrotron Source, Cornell University, Ithaca, NY, USA. ¹²Present address: Department of Electrical Engineering, Inha University, Incheon, South Korea. ¹³These authors contributed equally: Suraj S. Cheema, Daewoong Kwon. ✉e-mail: s.cheema@berkeley.edu; sayeef@berkeley.edu

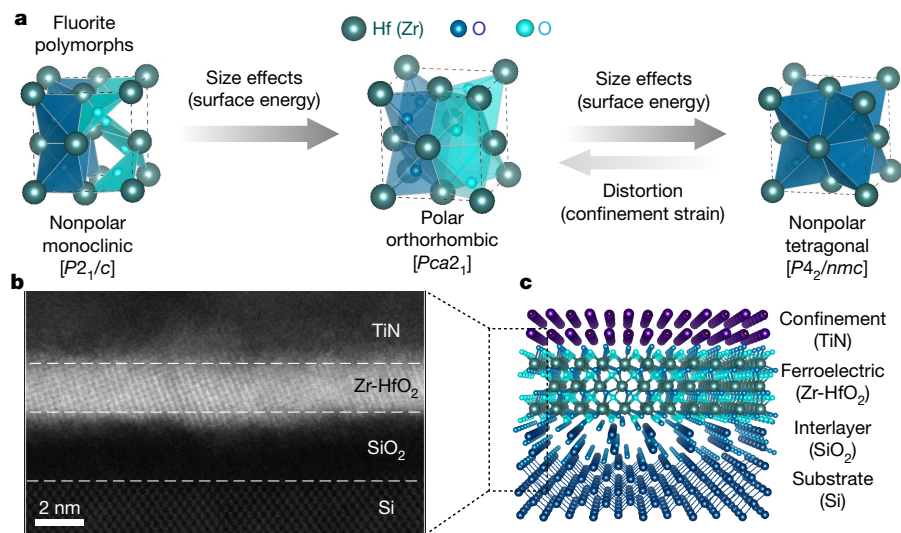


Fig. 1 | Size effects in fluorite-structure ferroelectrics. **a**, In fluorite-structure ferroelectrics, the polar distortion present in the orthorhombic phase can be represented as the centre anion displacement (cyan) with respect to its surrounding cation tetrahedron (represented by cyan arrow); in the nonpolar tetragonal phase, the oxygen atom (blue) lies in the polyhedral centre of the tetrahedron. The evolution of the bulk-stable monoclinic polymorph to the high-symmetry tetragonal and polar orthorhombic phases in the fluorite-structure structure illustrates the role of size effects – surface energies favour higher symmetry – and confinement strain – distortions favour lower symmetry – on stabilizing inversion asymmetry. In fluorite-structures, the noncentrosymmetric O-phase is higher

symmetry relative to the bulk-stable centrosymmetric M-phase. Consequently, both intrinsic (surface energy) and extrinsic (confinement strain) mechanisms can favour ultrathin inversion symmetry breaking in fluorite-structures, in stark contrast to size effect trends perovskites (Extended Data Fig. 1). **b**, Cross-sectional ADF STEM image of 20-cycle (about 1.8 nm) HZO, demonstrating ultrathin HZO films on silicon via low-temperature ALD. The Si substrate is oriented along the [110] zone axis. **c**, Schematic heterostructure investigated in this work, detailing the ultrathin ferroelectric HZO layer deposited on Si/SiO₂, and the capping metal layer employed to impart confinement strain during post-deposition rapid thermal annealing (Methods).

classical perovskites, surface-energy-driven size effects at reduced dimensions favour the high-symmetry paraelectric phase (cubic) over the low-symmetry ferroelectric phase (tetragonal)²⁰. Conversely, in fluorites, the noncentrosymmetric phase (orthorhombic *Pca*2₁, O-phase) is higher-symmetry relative to the bulk-stable centrosymmetric phase (*P*2₁/*c*, M-phase)²¹. Consequently, surface energies can promote (destabilize) inversion symmetry breaking in fluorite (perovskite) ferroelectrics in the two-dimensional limit^{21,22}. Owing to this size-induced noncentrosymmetry—that is, ‘reverse’ size effects (Fig. 1a, Extended Data Fig. 1)—HZO presents a promising model system in which to explore the ultrathin limits of ferroelectricity.

Thin films of HZO are grown using ALD down to ten cycles on oxidized silicon at 250 °C. For reference, approximately 11 ALD cycles correspond to 1 nm thickness, as confirmed by X-ray reflectivity (XRR, Extended Data Fig. 2) and transmission electron microscopy (TEM, Extended Data Fig. 3). Subsequently, HZO films are capped by a metal and subjected to rapid thermal annealing—henceforth referred to as confinement strain (Fig. 1c)—to impart anisotropic thermal stresses. Confinement strain has been reported to help distort the high-symmetry tetragonal fluorite-structure polymorph (*P*4₂/*nmc*, T-phase) into the polar O-phase¹³ (Fig. 1a). For confined ultrathin HZO films, X-ray diffraction analysis indicates a highly oriented noncentrosymmetric structure in contrast to polycrystalline thicker films (Extended Data Fig. 4). Inversion symmetry breaking in 10 cycle (about 1 nm) HZO films is confirmed by the presence of second harmonic generation (SHG) (Extended Data Fig. 5), previously employed to demonstrate ferroelectricity in a two-dimensional material²³.

Beyond inversion asymmetry, ferroelectricity also requires electrical switching between polarization states. Resonance-enhanced piezoresponse force microscopy (PFM) demonstrates bistable switching in Si/SiO₂/HZO heterostructures (Fig. 2a). PFM phase images on ten-cycle HZO films (Fig. 2d) show well defined regions of 180° phase contrast—corresponding to remanent polarization states (Fig. 2b)—that can be

rewritten in a non-volatile fashion. Notably, unpoled regions demonstrate the same phase contrast as positively poled regions (Fig. 2d), indicating that ultrathin HZO exhibits spontaneous polarization. In previous studies, field cycling is often required to ‘wake up’ the ferroelectricity in HfO₂-based fluorites, which is attributed to a field-induced nonpolar-polar phase transition¹³. Here, atomic-scale thickness in tandem with mechanical confinement enhances the polar phase stability to exhibit spontaneous polarization, eliminating one of the most critical issues plaguing fluorite-structure ferroelectrics¹⁶. Rapid thermal annealing alone is insufficient for ferroelectricity; regions of HZO annealed without a metal capping layer do not exhibit ferroelectric signatures (Extended Data Fig. 6). This highlights the critical role of ultrathin confinement, and the strains imposed by such layering, on stabilizing the polar O-phase in ultrathin fluorite-structure films.

Along with phase contrast imaging, local PFM switching spectroscopy further confirms the robust ferroelectricity in ten-cycle HZO films, as demonstrated by 180° phase hysteresis and its butterfly-shape amplitude (*d*₃₃) loops (Fig. 2e). PFM spectroscopy was performed on metal electrodes to eliminate electrostatic artefacts from the tip²⁴ and potential electromechanical contributions (Methods). In addition, careful monitoring of topography was performed during poling (Extended Data Fig. 7a) to detect the possibility of any electrochemical and electromechanical artefacts. Time-dependent PFM imaging (Extended Data Fig. 7b) on ten-cycle HZO illustrates polarization patterns sustained for at least 24 h; such long-term retention suggests that the PFM contrast is due to ferroelectric behaviour and not due to shorter-scale spurious effects often attributed to amorphous hafnia²⁵. Furthermore, alternating voltage (*V*_{ac})-dependent piezoresponse loops (Extended Data Fig. 7c) rule out electrostatic artefacts from charging²⁵. Moving beyond the standard PFM optical beam detection method, interferometric displacement sensor (IDS) PFM measurements (Extended Data Fig. 7d) definitively demonstrate the ferroelectric origin of switching spectroscopy hysteresis in ten-cycle HZO. The recently developed IDS

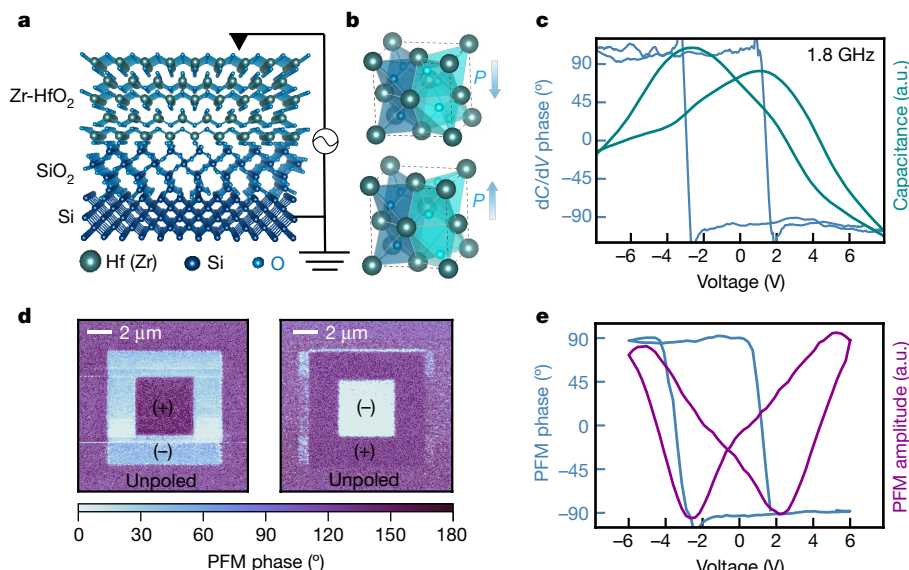


Fig. 2 | Electric polarization switching in ultrathin HZO. **a**, Schematic of the Si/SiO₂ (2 nm)/HZO (1 nm) heterostructure investigated by scanning probe imaging. **b**, Schematic of the HZO unit cell in the ferroelectric orthorhombic structure (*Pca*₂₁). The different-coloured oxygen atoms represent the displaced oxygen atoms (cyan) and the centrosymmetric oxygen atoms (blue) within the surrounding cation tetrahedron. The blue arrows labelled *P* denote the polarization directions corresponding to the acentric oxygen atomic displacements. **c**, Microwave-frequency SCM spectroscopy for a ten-cycle HZO film. The presence of butterfly-shaped *C*–*V* conclusively demonstrates ferroelectricity in ultrathin HZO, enabled by the high-frequency detection of differential capacitance (Methods). Microwave *dC/dV* measurements on multiple regions of ten-cycle HZO demonstrate the robust ferroelectric behaviour (Extended Data Fig. 8). **d**, Phase-contrast PFM images demonstrating stable, bipolar, remanent polarization states that can be

overwritten into the opposite polarization state for a ten-cycle HZO film. We note that the unpoled outer perimeter matches phase contrast with the positively poled regime regardless of the poling-polarity sequence; this indicates that ultrathin HZO exhibits spontaneous polarization without requiring ‘wake-up’ effects to become ferroelectric. Time-dependent PFM imaging further demonstrates the robust ferroelectric contrast (Extended Data Fig. 7). **e**, Phase and amplitude switching spectroscopy loops for a ten-cycle HZO film, demonstrating ferroelectric-like hysteresis. Interferometry-based IDS PFM hysteresis loops confirm that the origin of switching spectroscopy hysteresis is free of artefacts (Extended Data Fig. 7) and switching-spectroscopy measurements demonstrating the critical role of confinement during phase annealing for stabilizing the polar phase in ultrathin HZO (Extended Data Fig. 6).

technique²⁶ eliminates the long-range electrostatics and cantilever resonance artefacts that obfuscate typical voltage-modulated PFM. Contact IDS measurements demonstrate 180° phase hysteresis and butterfly-shaped *d*₃₃, which is indicative of ferroelectric behaviour free of electrostatic contributions²⁶. A lack of electrostatically driven hysteresis is confirmed by off-surface measurements (Extended Data Fig. 7d). Along with IDS, scanning capacitance microscopy (SCM) provides another advanced scanning probe technique with which to probe ferroelectricity in ultrathin HZO. SCM differential capacitance spectroscopy on ten-cycle HZO demonstrates butterfly-shaped capacitance–voltage (*C*–*V*) hysteresis (Fig. 2c, Extended Data Fig. 8). The microwave-frequency detection in SCM (Methods) mitigates the leakage currents that prevent typical bulk electrical characterization of ultrathin ferroelectrics, and provides conclusive evidence of ferroelectric polarization switching.

To examine the structural and electronic origins of ferroelectricity in HZO, we employ grazing-incidence X-ray diffraction (GI-XRD) and X-ray absorption spectroscopy (XAS) (Fig. 3). GI-XRD alone cannot unambiguously distinguish between certain fluorite-structure polymorphs in ultrathin HZO; as a complement to diffraction, XAS provides spectroscopic signatures of the polar O-phase and nonpolar T-phase. In particular, the T-phase nonpolar distortion (*D*_{4h}, fourfold prismatic symmetry) from regular tetrahedral to fluorite-structure symmetry does not split the degenerate *e* orbitals (*d*_{x²–y², *d*_{3z²–r²). Meanwhile, the O-phase polar rhombic pyramidal distortion (*C*_{2v}, twofold pyramidal symmetry) does split the *e* manifold, providing a symmetry-specific spectroscopic marker (Extended Data Fig. 9a,b). Owing to the *d*⁰ electronic configuration present in Hf⁴⁺ (Zr⁴⁺), spectral weight from oxygen *K*-edge XAS can be attributed solely to crystal field effects, providing}}

additional insights into the degree of structural distortion. The spectral weight of the symmetry-split *e* regimes (Extended Data Fig. 9c) and pre-edge regime (Extended Data Fig. 9e) increases with decreasing thickness, indicative of more pronounced rhombic distortion and divergence from isotropic nearest-neighbour oxygen polyhedral coordination (Extended Data Fig. 9d), respectively. These spectral weight trends provide further evidence of ultrathin-enhanced distortions. In conjunction with XAS, X-ray linear dichroism (XLD) can also probe structural distortions owing to its sensitivity to orbital asymmetry. Nanospectroscopy via soft X-ray photoemission electron microscopy (PEEM) illustrates spatially resolved XLD contrast at 535 eV (Extended Data Fig. 9f), corresponding to the *e*-split rhombic distortion regime. This suggests that XLD at the O *K* edge is indeed sensitive to polar features in ultrathin HZO. Shifting to sample-averaged XLD at the Zr *M*₂ edge, the orbital polarization is found to increase from the thick (100-cycle) to ultrathin (ten-cycle) regime (Fig. 3c), indicative of increased oxygen polyhedral distortion (Fig. 3d) consistent with ultrathin-enhanced ferroelectricity.

Remarkably, we also observe the emergence of crystallographic texturing of HZO films in the ultrathin regime (Fig. 1b, 3e). We note that many of the reflections in 100-cycle HZO, including the dominant (111), are absent in the GI-XRD spectra below 25 cycles owing to the geometric limitations of one-dimensional spectra (unable to detect all the reflections present in highly oriented films) (Methods). Tilted-geometry (*φ*–*χ*) diffraction (that is, pole figures) are required to access these oriented reflections at specific points, rather than polycrystalline-like rings, in reciprocal space. The spot-like patterns present in pole figures about the (111) reflections (Extended Data Fig. 4b) confirm the high degree of texturing in ultrathin HZO. Interestingly, this texturing

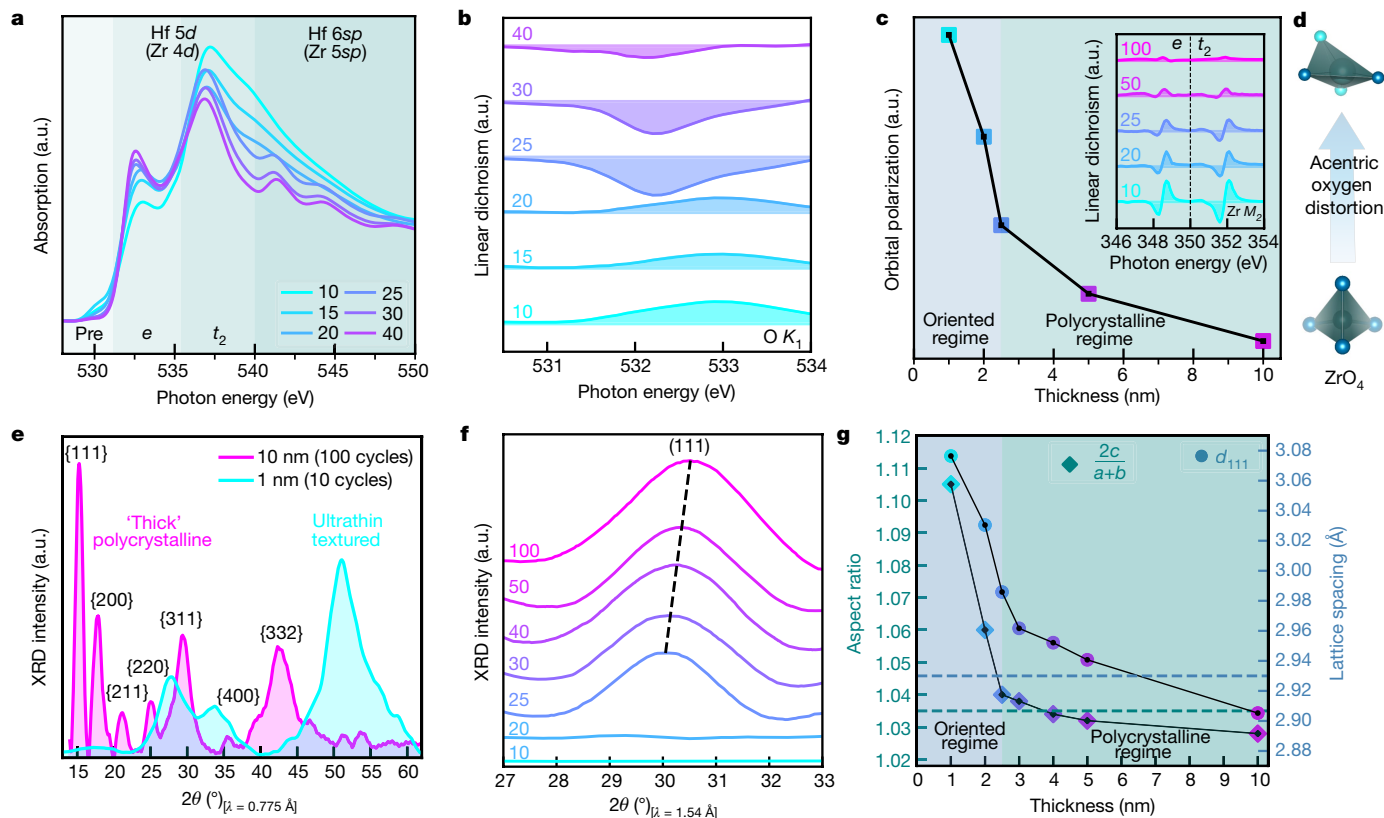


Fig. 3 | Emergence of 'reverse' size effects in ultrathin HZO. a,

Thickness-dependent XAS at the O K edge; spectral weight XAS trends indicate enhanced polyhedral disorder and tetrahedral and rhombic distortions in ultrathin films, illustrated by the crystal field splitting diagram for the fluorite-structure structural polymorphs and symmetry-specific XAS simulations (Extended Data Fig. 9a, b). The e and t_2 d -electron energy manifolds are set by the fluorite-structure tetrahedral symmetry (Extended Data Fig. 9a). **b,** Thickness-dependent XLD at the O K edge; the orbital polarization inversion below 25 cycles corresponds to the onset of highly oriented ultrathin films. **c,** Thickness-dependent orbital polarization and XLD (inset) at the Zr M_2 edge. The orbital polarization trend indicates ultrathin-enhanced ZrO_4 tetrahedral distortion, schematically represented in **d** by acentric oxygen atomic

displacement (cyan atoms). **e,** Synchrotron GI-XRD demonstrating the emergence of highly oriented ultrathin films, consistent with ADF-STEM (Fig. 1b) and pole figure analysis of ultrathin HZO (Extended Data Fig. 4b). **f,** Thickness-dependent GI-XRD around the polar orthorhombic (111) reflection, demonstrating a systematic shift in $2\theta_{111}$ with thickness, and highlighting the limitation of GI-XRD geometry to detect the (111)-reflection below 25 cycles as the film becomes highly oriented (Extended Data Fig. 4). **g,** Thickness-dependent d_{111} lattice spacing and $2c/(a+b)$ structural aspect ratio, suggesting amplified polarization in the ultrathin limit, especially below 25 cycles. Dashed lines denote reported d_{111} and aspect ratio values for thicker ferroelectric HZO films. Aspect ratio values are extracted from the symmetry-split {200} planes (Methods).

happens despite local nanocrystalline regions observed in TEM for ultrathin films (for example, 15-cycle HZO in Extended Data Fig. 3d). Coinciding with the onset of texturing, the microstructural evolution below 25-cycle HZO manifests spectroscopically as inverted orbital polarization at the e manifold (Fig. 3b), suggesting flipped polar-distortion-split e levels ($d_{x^2-y^2}$ and $d_{3z^2-r^2}$). This indicates that sub-25-cycle ultrathin films enter a new electronic structure concurrently as the crystalline structure orders. Therefore, confinement strain in atomic-scale fluorite films could provide a route to tailor electronic structure and engineer polarization at the orbital level²⁷, akin to epitaxial strain in perovskite films.

The onset of highly ordered films also coincides with sharp rises in structural markers of distortion (Fig. 3g). The degree of rhombic distortion is captured by the lattice spacing d_{111} ; accordingly, d_{111} is tied to macroscopic polarization in HZO^{28,29}. We observe increasing d_{111} with decreasing thickness (Fig. 3g), as previously reported in epitaxial HZO films grown by pulsed laser deposition^{28,29}, consistent with ultrathin-enhanced ferroelectricity. Notably, our low-temperature ALD-grown films on silicon can induce similar structurally induced phenomena observed in high-temperature pulsed-laser-deposition-grown epitaxial films on perovskite templates and extended to an even thinner limit. Furthermore, the d_{111} bifurcation below 25 cycles suggests a link between texturing and amplified distortion in the ultrathin

regime. Another crystallographic signature, orthorhombic aspect ratio ($2c/(a+b)$), also indicates enhanced distortions in the ultrathin regime (Fig. 3g). Fluorite-structure orthorhombicity¹³ is akin to perovskite tetragonality³⁰ (c/a); these ratios serve as structural barometers of macroscopic polarization. The orthorhombic distortion present in ten-cycle HZO far exceeds any reported values for HfO_2 - ZrO_2 polymorphs¹³: we find >10% aspect asymmetry, whereas 3–4% is typically reported for fluorite-structure ferroelectrics, consistent with our thicker films (Fig. 3g). Correspondingly, the tetrahedral and rhombic crystal field splitting energies in ultrathin films surpass expected polar fluorite-structure values by 1.3 eV and 700 meV, respectively (Extended Data Fig. 9g). Such colossal structural splittings are well beyond the reported limits of epitaxial strain in perovskite films²⁷. Therefore, although prohibitive tunnel currents prevent accurate quantification of polarization from traditional polarization–voltage measurements, multiple structural gauges of polarization indicate substantial enhancement in the ultrathin limit.

In summary, several techniques self-consistently demonstrate robust ferroelectricity in HZO films of thickness down to 1 nm (Extended Data Fig. 2, 3), synthesized by low-temperature ALD on silicon. Remarkably, these experiments indicate that polar distortions are amplified in the ultrathin limit; diffraction markers (d_{111} lattice spacing, structural aspect ratio) and spectroscopic signatures (orbital polarization, crystal

field splitting) all demonstrate ultrathin enhancement. Such ‘reverse’ size effects oppose conventional perovskite ferroelectric trends⁶. Previous works on polycrystalline doped HfO_2 ^{31–33} have explained thickness-dependent polarization trends based on the volume fraction of the ferroelectric O-phase. However, our observations are more consistent with studies of pseudo-epitaxial HZO films grown on perovskite templates^{28,29} in that we also observe substantial orientation in the ultrathin regime and enhancement of ferroelectricity with decreasing thickness. Importantly, our work demonstrates that this enhancement persists down to at least two fluorite-structure unit cell thickness, overcoming the deleterious depolarization field effects that would otherwise dominate a prototypical perovskite-structure ferroelectric in this ultrathin regime^{1,34}. Further studies should explore how the current understanding of film synthesis, phase competition and polar distortion in HZO developed for thicker films ($>5\text{ nm}$)³⁵ evolves in the ultrathin regime ($<2\text{ nm}$). Our results indicate that harnessing confinement strain to amplify atomic displacements in ultrathin films provides a route towards enhancing electric polarization at the nanoscale beyond epitaxial strain^{36,37}, akin to strain gradients in flexoelectricity^{38,39}. From a technological perspective, direct monolithic integration of ultrathin doped HfO_2 on Si/SiO₂ paves the way for polarization-driven low-power memories (Extended Data Fig. 10) and ultra-scaled ferroelectrics-based transistors^{40,41}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at [<https://doi.org/10.1038/s41586-020-2208-x>].

- Junquera, J. & Ghosez, P. Critical thickness for ferroelectricity in perovskite ultrathin films. *Nature* **422**, 506–509 (2003).
- Mikolajick, T., Slesazeck, S., Park, M. & Schroeder, U. Ferroelectric hafnium oxide for ferroelectric random-access memories and ferroelectric field-effect transistors. *MRS Bull.* **43**, 340–346 (2018).
- Dawber, M., Rabe, K. M. & Scott, J. F. Physics of thin-film ferroelectric oxides. *Rev. Mod. Phys.* **77**, 1083–1130 (2005).
- Schlom, D. G., Guha, S. & Datta, S. Gate oxides beyond SiO₂. *MRS Bull.* **33**, 1017–1025 (2008).
- Lines, M. E. & Glass, A. M. *Principles and Applications of Ferroelectrics and Related Materials* (Oxford Univ. Press, 1977).
- Ahn, C., Rabe, K. & Triscone, J. Ferroelectricity at the nanoscale: local polarization in oxide thin films and heterostructures. *Science* **303**, 488–491 (2004).
- Fong, D. D. et al. Ferroelectricity in ultrathin perovskite films. *Science* **304**, 1650–1653 (2004).
- Tenne, D. A. et al. Probing nanoscale ferroelectricity by ultraviolet Raman spectroscopy. *Science* **313**, 1614–1616 (2006).
- Warusawithana, M. P. et al. A ferroelectric oxide made directly on silicon. *Science* **324**, 367–370 (2009).
- Dubourdieu, C. et al. Switching of ferroelectric polarization in epitaxial BaTiO₃ films on silicon without a conducting bottom electrode. *Nat. Nanotechnol.* **8**, 748–754 (2013).
- Schlom, D. G. & Haeni, J. H. A thermodynamic approach to selecting alternative gate dielectrics. *MRS Bull.* **27**, 198–204 (2002).
- Böscke, T. S., Müller, J., Bräuhäus, D., Schröder, U. & Böttger, U. Ferroelectricity in hafnium oxide thin films. *Appl. Phys. Lett.* **99**, 102903 (2011).
- Park, M. H. et al. Ferroelectricity and antiferroelectricity of doped thin HfO_2 -based films. *Adv. Mater.* **27**, 1811–1831 (2015).
- Robertson, J. High dielectric constant gate oxides for metal oxide Si transistors. *Rep. Prog. Phys.* **69**, 327 (2006).
- Muller, J. et al. Ferroelectric hafnium oxide: a CMOS-compatible and highly scalable approach to future ferroelectric memories. In *2013 IEEE Int. Electron Devices Meet. (IEDM)* 10.8.1–10.8.4 (IEEE, 2013).
- Park, M., Lee, Y., Mikolajick, T., Schroeder, U. & Hwang, C. Review and perspective on ferroelectric HfO_2 -based thin films for memory applications. *MRS Commun.* **8**, 795–808 (2018).
- Wong, J. C. & Salahuddin, S. Negative capacitance transistors. *Proc. IEEE* **107**, 49–62 (2019).
- Kwon, D. et al. Improved subthreshold swing and short channel effect in FDSOI n-channel negative capacitance field effect transistors. *IEEE Electron Device Lett.* **39**, 300–303 (2018).
- Salahuddin, S., Ni, K. & Datta, S. The era of hyper-scaling in electronics. *Nat. Electron.* **1**, 442–450 (2018).
- Merz, W. J. The effect of hydrostatic pressure on the Curie point of barium titanate single crystals. *Phys. Rev.* **78**, 52 (1950).
- Ohtaka, O. et al. Phase relations and volume changes of hafnia under high pressure and high temperature. *J. Am. Ceram. Soc.* **84**, 1369–1373 (2001).
- Materlik, R., Künne, C. & Kersch, A. The origin of ferroelectricity in $\text{Hf}_{1-x}\text{Zr}_x\text{O}_2$: a computational investigation and a surface energy model. *J. Appl. Phys.* **117**, 134109 (2015).
- Xiao, J. et al. Intrinsic two-dimensional ferroelectricity with dipole locking. *Phys. Rev. Lett.* **120**, 227601 (2018).
- Vasudevan, R. K., Balke, N., Maksymovych, P., Jesse, S. & Kalinin, S. V. Ferroelectric or non-ferroelectric: why so many materials exhibit “ferroelectricity” on the nanoscale. *Appl. Phys. Rev.* **4**, 021302 (2017).
- Balke, N. et al. Differentiating ferroelectric and nonferroelectric electromechanical effects with scanning probe microscopy. *ACS Nano* **9**, 6484–6492 (2015).
- Collins, L., Liu, Y., Ovchinnikova, O. S. & Proksch, R. Quantitative electromechanical atomic force microscopy. *ACS Nano* **13**, 8055–8066 (2019).
- Disa, A. S. et al. Orbital engineering in symmetry-breaking polar heterostructures. *Phys. Rev. Lett.* **114**, 026801 (2015).
- Wei, Y. et al. A rhombohedral ferroelectric phase in epitaxially strained $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ thin films. *Nat. Mater.* **17**, 1095–1100 (2018).
- Lyu, J., Fina, I., Solanas, R., Fontcuberta, J. & Sánchez, F. Growth window of ferroelectric epitaxial $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ thin films. *ACS Appl. Electron. Mater.* **1**, 220–228 (2019).
- Schlom, D. G. et al. Elastic strain engineering of ferroic oxides. *MRS Bull.* **39**, 118–130 (2014).
- Park, M. H. et al. Evolution of phases and ferroelectric properties of thin $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ films according to the thickness and annealing temperature. *Appl. Phys. Lett.* **102**, 242905 (2013).
- Tian, X. et al. Evolution of ferroelectric HfO_2 in ultrathin region down to 3 nm. *Appl. Phys. Lett.* **112**, 102902 (2018).
- Richter, C. et al. Si doped hafnium oxide—a “fragile” ferroelectric system. *Adv. Electron. Mater.* **3**, 1700131 (2017).
- Stengel, M. & Spaldin, N. A. Origin of the dielectric dead layer in nanoscale capacitors. *Nature* **443**, 679–682 (2006).
- Kim, S. J., Mohan, J., Summerfelt, S. R. & Kim, J. Ferroelectric $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ thin films: a review of recent advances. *JOM* **71**, 246–255 (2019).
- Schlom, D. G. et al. Strain tuning of ferroelectric thin films. *Annu. Rev. Mater. Res.* **37**, 589–626 (2007).
- Haeni, J. H. et al. Room-temperature ferroelectricity in strained SrTiO_3 . *Nature* **430**, 758–761 (2004).
- Zubko, P., Catalan, G. & Tagantsev, A. K. Flexoelectric effect in solids. *Annu. Rev. Mater. Res.* **43**, 387–421 (2013).
- Jariwala, D., Marks, T. J. & Hersam, M. C. Mixed-dimensional van der Waals heterostructures. *Nat. Mater.* **16**, 170–181 (2017).
- Kwon, D. et al. Negative capacitance FET with 1.8-nm-thick Zr-doped HfO_2 oxide. *IEEE Electron Device Lett.* **40**, 993–996 (2019).
- Lee, M. H. et al. Physical thickness 1.x nm ferroelectric HfZrO_x negative capacitance FETs. In *2016 IEEE Int. Electron Devices Meet. (IEDM)* 12.1.1–12.1.4, <https://ieeexplore.ieee.org/document/7838400/> (IEEE, 2016).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Sample deposition and preparation

Thin films of $\text{Hf}_{0.8}\text{Zr}_{0.2}\text{O}_2$ were grown by ALD in a Fiji Ultratech/Cambridge Nanotech tool at 250 °C, in which tetrakis (ethylmeth-ylamino) hafnium and tetrakis (ethylmeth-ylamino) zirconium pre-cursors are heated to 75 °C and water vapour is used as the oxidant. For metal-ferroelectric-insulator-semiconductor structures, heavily p-doped Si(100) substrates (10^{19} cm^{-3}) are first oxidized in ambient O_2 during an rapid thermal annealing step at 900 °C for 60 s, forming about 2 nm of thermal SiO_2 on Si. For metal-ferroelectric-metal struc-tures, the Si substrate is coated with 30 nm TiN. Subsequently, HZO is deposited at 250 °C by ALD; a 4:1 ratio between the HfO_2 monolayer and the ZrO_2 monolayer sets the 80:20 stoichiometry of the depos-ited HZO, in which ten cycles corresponds to 1 nm of film. After ALD deposition, a top metal (W or TiN) is deposited by sputtering at room temperature. Finally, a rapid post-metal annealing at 500 °C (30 s, ambient N_2 background) stabilizes the desired polar orthorhombic phase. For capacitor structures (scanning probe studies), the top electrodes are defined by photolithography and dry etching. For bare structures (structural studies), the top metal is removed by chemical etching to expose the HZO surface. Further details pertaining to ALD growth conditions, post-deposition processing, and so on are outlined in a previous work⁴². All thin film synthesis was performed at Univer-sity of California, Berkeley; processing was performed at the Marvell Nanofabrication Laboratory at University of California, Berkeley. One nanometre of chemically grown SiO_2 on Si was prepared by the standard clean (SC-1) solution (5:1:1 $\text{H}_2\text{O}:\text{H}_2\text{O}_2:\text{NH}_4\text{OH}$ at 80 °C for 10 min) after the Si wafer was cleaned in Piranha (120 °C for 10 min) to remove organ-ics and HF (50:1 $\text{H}_2\text{O}:\text{HF}$ at room temperature for 30 s) to remove any native oxide. Thinner SiO_2 was employed to help reduce depolarization fields and improve the electric field distribution through the ultrathin ferroelectric HZO layer.

Electron microscopy

Electron microscopy was performed at the National Center for Electron Microscopy facility of the Molecular Foundry at LBNL as well as by Nanolab Technologies Inc., a commercial vendor. At the National Center for Electron Microscopy, TEM samples were prepared by mechanical polishing on an Allied High Tech Multiprep and subsequently Ar ion milled using a Gatan Precision Ion Milling System at shallow angles (5° to 3°) with starting energies of 5 keV stepped down to a final cleaning energy of 200 eV to reduce ion-induced damage. High-angle annu-lar dark-field (HAADF) scanning transmission electron microscopy (STEM) images were recorded on TEAMI, an aberration-corrected FEI Titan 80–300 operated in STEM mode at 300 kV with a convergence semi-angle of 17 mrad, 70 pA probe current, and collection angles >40 mrad. The local thicknesses of the respective HZO layers were determined from calibration to the Si(110) interplanar lattice spacing (Extended Data Fig. 3), consistent with global thicknesses extracted using XRR (Extended Data Fig. 2).

Scanning probe microscopy

Piezoresponse microscopy and spectroscopy. PFM measurements (Extended Data Figs. 6, 7) were performed using a commercial scanning probe microscope (Asylum MFP-3D) at the University of California, Berkeley. Dual-frequency resonance-tracking PFM⁴³ was conducted using a conductive Pt/Ir-coated probe tip (NanoSensor PPP-EFM) to image written domain structures and measure switching-spectroscopy⁴⁴ piezoelectric hysteresis loops. Resonance-enhanced PFM increases the signal-to-noise ratio for the detection of out-of-plane electric po-larization, which is critical for ultrathin films. Contact was made to the bottom TiN electrode or heavily doped Si substrate for grounding in PFM studies. All PFM phase-contrast images and hysteresis loops shown were performed on ten-cycle (about 1 nm) HZO films unless

otherwise indicated. PFM imaging was performed with the tip in direct contact with the HZO layer. Switching-spectroscopy hysteresis loops were measured on capacitor structures to help eliminate electrostatic artefacts from the tip⁴⁵, mitigate possible electromechanical contri-butions²⁴, and to yield more confined electric fields. V_{ac} -dependent piezoresponse loops (Extended Data Fig. 7c) examined the ferroelectric origin of the PFM signal—as opposed to tip bias-induced artefacts²⁵—in ultrathin HZO films. The piezoresponse OFF loop collapsed once V_{ac} exceeded the coercive voltage⁴⁶, as expected for ferroelectric behav-iour. Piezoresponse is defined as $A\cos\theta$, where A and θ are the PFM amplitude and phase, respectively⁴⁵. We note that the non-ideal shape of the piezoresponse loops, particularly at higher voltages, are caused by non-ferroelectric artefacts from the additional dielectric SiO_2 layer through which most of the voltage is dropped. For all PFM studies, the bias was applied to the tip.

Effective coercive field. The switching voltage from PFM loops exaggerate the coercive field of the ultrathin HZO layer once considering the potential distribution across the modified metal-oxide-semiconductor structure (oxide bilayer SiO_2 -HZO). The effective co-ercive field of the HZO layer can be determined using a simple dielectric-ferroelectric bilayer model, ignoring accumulation and depletion regions at the moment just to approximate the coercive field. Considering appropriate electrical boundary conditions across the oxide interface ($\epsilon_{DE}E_{DE} = \epsilon_{FE}E_{FE}$), the voltage across the ferroelectric layer (V_{FE}) can be expressed in terms of the total voltage given by the PFM loop (V_{tot}):

$$V_{FE} = \left(1 + \frac{t_{DE} \epsilon_{FE}}{t_{FE} \epsilon_{DE}}\right)^{-1} V_{tot}$$

where the dielectric constant for the oxide layers are taken as $\epsilon_{DE} = 3.9$ (SiO_2) and as $\epsilon_{FE} = 24$ (HZO)^{4,14} and the thicknesses of the oxide layers are $t_{DE} = 2$ nm and $t_{FE} = 1$ nm. These values yield $V_{FE} = V_{tot}/13$, so the effective coercive field of the ten-cycle (1 nm) ferroelectric HZO layer is approximately 2 MV cm^{-1} , consistent with values of thicker HZO films reported in the literature¹³.

Interferometry. IDS PFM measurements (Extended Data Fig. 7d) were performed using a commercial scanning probe microscope (Asylum Cypher) with an integrated quantitative laser Doppler vibrometer at Asylum Research (Santa Barbara). This recently developed method²⁶ eliminates crosstalk and other artefacts present in voltage-modulated piezo-measurements (d_{33}) by replacing the typical slope-sensitive optical beam detector with a displacement sensitive interferometer. By positioning the IDS laser directly over the tip, motion of the tip can be decoupled from spurious motion of the cantilever body. Motion of the cantilever body can be driven by long-range electrostatics and is influenced by the transfer function of the cantilever²⁶. IDS measure-ments were performed with a 3 N m^{-1} Ti/Ir coated cantilever placed on the bare 1-nm HZO surface at drive frequency 250 kHz and average force 75 nN. Off-surface loops (tip raised from surface), which measure the extrinsic electrostatic contributions²⁶, were performed by chang-ing the trigger value from deflection (force) to the z-sensor read-out. The lack of hysteresis from off-surface (non-contact) loops further support the finding that the hysteresis observed from on-contact IDS measurements are free from electrostatic contributions. Typical voltage-modulated PFM measurements often display false ferroelectric hysteresis due to long-range cantilever dynamics inherent to detection by the optical beam detector, as observed from non-contact hysteresis in non-piezoelectric samples²⁶.

Microwave capacitance. SCM measurements (Extended Data Fig. 8c) were performed using a commercial scanning probe microscope (Asylum Cypher) at Asylum Research (Santa Barbara). Differential

Article

capacitance (dC/dV) measurements were performed at 1.8-GHz frequency with a 40 kHz lock-in frequency at 0.5 V_{ac}. Pure Pt cantilevers (Rocky Mountain Nanotechnology) are placed on top of bare HZO surface (contact mode) on TiN-buffered Si for SCM measurements; dC/dV signals were extracted via V_{ac} applied between the SCM tip and bottom electrode. Capacitance–voltage loops via SCM have been previously used to confirm ferroelectricity in SrBi₂Ta₂O₉ (SBT) thin films⁴⁷. The microwave-frequency nature of the measurement lends itself to probing ultrathin ferroelectrics, as it mitigates leakage contributions. For SCM measurements, the bias was applied to the sample (swept up to ±8 V), not to the tip as is done for PFM measurements.

X-ray diffraction

Structural characterization. Synchrotron GI-XRD (Extended Data Fig. 4a) was performed at the Sector 33-BM-C beamline of the Advanced Photon Source, Argonne National Laboratory. Using synchrotron GI-XRD, we investigated the structural evolution from polycrystalline bulk-like (100-cycle) HZO down to highly textured ultrathin (<25 cycles) HZO in its polar orthorhombic (*Pca*2₁) phase, at grazing angle $\leq \theta = 0.35^\circ$. The high flux from the synchrotron source ($\lambda = 0.775 \text{ \AA}$) enabled collection of sufficient diffraction intensity from the few crystallographic planes present in ultrathin HZO samples. High-resolution GI-XRD was also performed using a laboratory-based Panalytical X'Pert Pro X-ray diffraction system (Cu K_α radiation, $\lambda = 1.54056 \text{ \AA}$) on HZO films thicker than 2 nm at grazing angle $\theta = 0.35^\circ$. Previous work employed selected area electron diffraction⁴⁸ and convergent beam electron diffraction⁴⁹ to attribute ferroelectricity in HfO₂-based films to the polar orthorhombic (*Pca*2₁) phase. The indexing of ultrathin HZO films performed in this work is consistent with the same polar orthorhombic phase determined from these previous electron diffraction studies.

Texture analysis. Pole figures (Extended Data Fig. 4b) were measured at Sector 33-BM-C beamline of the Advanced Photon Source, Argonne National Laboratory. For fixed *Q* values—corresponding to the *d*₁₁₁ lattice spacing—the 4-circle Huber diffractometer rotated in-plane (ϕ) 360° at multiple values of out-of-plane tilt (χ). The PILATUS 100K pixel area detector collected volumetric reciprocal space data from which two-dimensional pole figure slices were plotted for shells of constant *Q*_z. The four concentrated reflections for the {111} projection in *Q*_x–*Q*_y space indicate highly oriented texture, rather than the diffuse rings expected for polycrystalline films. As indicated in the main text (Fig. 3), films thinner than 25 cycles display substantial texturing (Extended Data Fig. 4); in particular, the (111) reflection, which is dominant for thicker films, is diminished in GI-XRD spectra of ultrathin films owing to the geometric limitations of a one-dimensional pattern (it is unable to detect all reflections present in oriented films). This limitation necessitates tilted-geometry two-dimensional patterns (pole figures) in order to detect all reflections present in highly oriented films. Meanwhile, reflections corresponding to polycrystalline films exhibit continuous rings in two-dimensional *Q*_x–*Q*_y reciprocal space, so any one-dimensional line-cut (GI-XRD spectra) would detect all reflections present. The results on ultrathin HZO films are in stark contrast to results on thicker films^{50,51} and indicate that, for such ultrathin films, crystallization and orientation need to be considered together. We also observed that for such thin films, the template for HZO growth needs to be atomically smooth. Therefore, the Si/SiO₂ interface employed in this work is critical; the growth surface is expected to have a larger role for ultrathin films than for thicker films.

Thickness confirmation. Synchrotron XRR of ultrathin HZO films (Extended Data Fig. 2)—performed at Sector 33-BM-C beamline of the Advanced Photon Source, Argonne National Laboratory and at Beamline 2-1 of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory—confirmed the thickness of sub-2-nm films. Fitting analysis was performed with the Python package

xrayutilities⁵². XRR of thicker HZO films (>2 nm) was measured with the Panalytical X'Pert Pro system, and thickness fitting was performed with Panalytical software. The extracted growth rate of 11 cycles nm^{−1} is consistent with results from TEM and literature¹³.

Structural distortion analysis. For the polar orthorhombic phase (*Pca*2₁), we consider the orthorhombic distortion (that is, orthorhombicity) as the aspect ratio: $2c/(a+b)$ to enable easier comparison to the tetragonal (*P4*₂/nm) aspect ratio c/a . Fluorite-structure orthorhombicity is meant to be analogous to the perovskite ferroelectric tetragonality (c/a , where c is the polar axis); both aspect ratios serve as a structural gauge of the macroscopic polarization because they are indicative of the polar distortion present in their respective structures⁵³. Notably, the orthorhombic distortion present in HZO is enhanced in the ultrathin regime—opposite to the typical tetragonal distortion trend in perovskite ferroelectrics⁶—indicative of the ‘reverse’ size effects present in fluorite-structure ferroelectrics. For example, the tetragonal aspect ratio was shown to decrease with decreasing thickness in ferroelectric PbTiO₃ films⁵³, while the orthorhombic aspect ratio is greatly enhanced in the ultrathin regime in our fluorite-structure HZO films (Fig. 3g). The orthorhombic distortion present in ten-cycle (about 1 nm) HZO far exceeds any reported values for HfO₂–ZrO₂ polymorphs: we find >11% aspect asymmetry, while 3–4% is typically reported for fluorite-structure ferroelectrics³¹, consistent with our thicker films (Fig. 3g). Indeed, a strong relationship between this aspect ratio and the remanent polarization value has been experimentally demonstrated in thicker doped HfO₂ films⁵⁴. Therefore, the colossal orthorhombic aspect ratio present in ten-cycle HZO is consistent with ultrathin-enhanced ferroelectricity. The orthorhombic aspect ratio is calculated from the position of various diffraction peaks indexed to the *Pca*2₁ phase (the 200, 020 and 002 peaks), using the following crystallographic relations: $a = 2 \cdot d_{200}$, $b = 2 \cdot d_{020}$, $c = 2 \cdot d_{002}$ where d_{200} , d_{020} , and d_{002} are the 200, 020 and 002 lattice spacings determined via Bragg's law and the respective peak positions. These values are self-consistently checked against the 111 interplanar lattice spacing ($1/a_{111}^2 = 1/a^2 + 1/b^2 + 1/c^2$) as well as against other orientations present in the diffraction spectra. The aspect ratio of the polar O-phase exceeds that of the T-phase (c/a) for doped HfO₂⁵⁴. Another structural marker indicates amplified distortions as thickness is reduced, namely the interplanar lattice spacing d_{111} . The origin of the left shift in the O-phase 111 (T-phase 101) reflection (Fig. 3f) with decreasing thickness (that is, decreasing ALD cycles) is typically attributed to the abovementioned phase transition (nonpolar T-phase to polar O-phase); the left shift of the peak in reciprocal space corresponds to an increase in real-space lattice spacing. Extending this analogy to the ultrathin regime in which the polar O-phase is already stabilized, the ultrathin enhancement of d_{111} (Fig. 3g) indicates a further increase in rhombic distortion (structurally represented by d_{111}). Recent works on epitaxial HZO films grown by high-temperature pulsed laser deposition on perovskite substrates also indicate increasing d_{111} with decreasing thickness^{28,29}; these works find the electric polarization to increase with increasing d_{111} . Similarly, we expect a larger polarization in our ultrathin films based on the d_{111} trend (Fig. 3g); notably, our low-temperature ALD-grown highly oriented films are mimicking the trends observed in high-temperature pulsed-laser-deposition-grown epitaxial films.

X-ray spectroscopy

XAS and XLD. X-ray absorption spectroscopy (XAS) and XLD was performed at the Advanced Light Source beamline 4.0.2. XAS measurements were taken at the oxygen *K* edge (520–550 eV) and Zr *M*₂ edge (345–355 eV). X-rays were incident at 20° off grazing. XLD (XAS) was obtained from the difference (average) of horizontal and vertical linearly polarized X-rays. To eliminate systematic artefacts in the signal that drift with time, spectra were captured with the order of polarization rotation reversed (such as horizontal, vertical, vertical and horizontal)

in successive scans. An elliptically polarizing undulator was used to tune polarization and photon energy of the synchrotron X-ray source⁵⁵. XAS was recorded under total electron yield mode⁵⁵.

Simulated XAS and crystal field symmetry. Simulated XAS spectra for the various fluorite-structure polymorphs were computed through the Materials Project⁵⁶ open-source database for the XAS spectrum⁵⁷. In particular, the following symmetries for HfO₂ and ZrO₂ were investigated: monoclinic $P2_1/c$ (space group 14), orthorhombic $Pca2_1$ (space group 29), and tetragonal $P4_2/nmc$ (space group 137). Comparisons between HZO and the undoped fluorite-structure endmembers (in particular, qualitative comparison of splitting-induced spectroscopy features) are reasonable owing to the extremely low structural dissimilarity between the same polymorphs of HfO₂ and ZrO₂, as determined by pymatgen⁵⁸. The T-phase ($P4_2/nmc$) nonpolar distortion (D_{4h} , fourfold prismatic symmetry) from regular tetrahedral (T_d , full tetrahedral symmetry) fluorite-structure symmetry does not split the degenerate e bands ($d_{x^2-y^2}$, $d_{3z^2-r^2}$), as confirmed by experiment⁵⁹ and the XAS simulations (Extended Data Fig. 9b). Meanwhile, the O-phase ($Pca2_1$) polar rhombic pyramidal distortion (C_{2v} , twofold pyramidal symmetry) does split the e -manifold based on crystal field symmetry (Extended Data Fig. 9b), providing a spectroscopic means to distinguish the T- and O-phases. The eightfold Hf-O (Zr-O) coordination (Extended Data Fig. 9d) in the tetragonal phase (D_{2d} point group symmetry) can be decomposed into two tetrahedra that are the space inversion twins of one another. Therefore, crystal field splitting of the e levels matches that of a single tetrahedron⁵⁹—that is, there is no further splitting. Meanwhile, the sevenfold Hf-O (Zr-O) coordination (Extended Data Fig. 9d) in the orthorhombic phase cannot be decomposed into two tetrahedra; the additional rhombic distortion (not present in the T-phase) splits the e manifold. The simulated XAS spectra for T- and O-phase ZrO₂ (Extended Data Fig. 9b) supports this picture, because the additional spectroscopic feature present between the main e - and t_2 -absorption features in the O-phase is presumably caused by this additional symmetry-lowering distortion. The XAS spectra of the HZO thickness series (Extended Data Fig. 9c) demonstrates tetrahedral and rhombic splitting features closely matching the polar O-phase ($Pca2_1$). This demonstrates a spectroscopic method for phase identification beyond diffraction—ambiguous owing to the nearly identical T- and O-phase lattice parameters¹³—whose signatures are more sensitive to the subtle structural distortions present as symmetry is lowered from the T- to the O-phase.

Crystal field splitting. Notably, the crystal field distortions present in confined HZO films greatly exceed what is typically observed in bulk fluorite-structures and perovskite ferroelectrics (Extended Data Fig. 9g); the tetrahedral (rhombic) crystal field Δ_T (Δ_R) arising from the T_d (C_{2v}) symmetry in ten-cycle HZO films is 1.3 eV (0.7 eV) greater than what is expected from fluorite-structure ZrO₂ in the polar orthorhombic phase ($Pca2_1$). The computational XAS for the $Pca2_1$ phase already takes the polar distortion (Δ_R) into account; so the enhanced Δ_R in ultrathin confined films again points to enhanced polar distortions (consistent with diffraction-based results). Kindred efforts to uncover routes towards enhanced nanoscale distortions have been explored in complex perovskite heterostructures. For example, in nickelate perovskite superlattices, enormous Δ_{eg} crystal field splitting (up to 0.8 eV) has been achieved via polar fields resulting from internal charge transfer²⁷; >10% epitaxial strain would be required to induced such large ionic distortions in that particular system, well beyond the limits of epitaxial strain, which can only achieve e_g splitting of about 300 meV (ref. ⁶⁰).

Spectral weight trends. The relative spectral weight of the e and t_2 manifolds (Extended Data Fig. 9c) at the O K edge can also provide insight into the degree of structural distortion⁶¹. Owing to the d^0 electronic configuration present in Hf⁴⁺ (Zr⁴⁺), all d states are available for

mixing with O $2p$ states, so the analysis of e - t_2 spectral weight can be simplified to be purely due to crystal field effects⁶¹. Tetrahedral symmetry lowers e bands relative to t_2 bands due to the enhanced t_2 orbital overlap with oxygen $2p$ orbitals. The enhanced t_2/e spectral weight as thickness is reduced (Extended Data Fig. 9c) indicates the preference for O $2p$ hybridization with Hf $5d$ (Zr $4d$) t_2 orbitals, further exaggerating the disparity set by the tetrahedral symmetry as the symmetry is lowered to the polar O-phase. Additionally, the increase in spectral weight of the pre-edge shoulder (Extended Data Fig. 9e) provides further confirmation that structural distortions are amplified in the ultrathin limit. Pre-edge features at the O K edge in complex transition metal oxides are commonly attributed to nearest-neighbour variations from typical oxygen polyhedral coordination as the symmetry is lowered by various distortions⁶². Analogously, here the pre-edge feature is attributed to variation from eightfold coordination in the T-phase (NN = 8) as the symmetry is lowered into the polar O-phase (NN = 7) (Extended Data Fig. 9d). On the unit cell level in the polar O-phase, the central metal cation is surrounded by an asymmetric oxygen coordination environment (note the 4 blue and 3 cyan oxygen atoms in Extended Data Fig. 9d) owing to the polar rhombic distortion of normal tetrahedral (T_d) symmetry; this polyhedral distortion can manifest as increased spectral weight at the oxygen K pre-edge⁶². The critical e manifold splitting due to the polar rhombic distortion also increases in spectral weight as thickness is reduced (Extended Data Fig. 9c). The XAS spectral weight trends mirror the structural indicators of ultrathin-enhanced distortion (Fig. 3c).

Orbital polarization. In conjunction with XAS, XLD can also probe structural distortions owing to its sensitivity to orbital asymmetry, which can arise from inversion symmetry breaking. For example, in the perovskite ferroelectrics PbTiO₃ and BaTiO₃, the Ti $3d$ to O $2p$ orbital hybridization is essential for stabilizing the noncentrosymmetric structure⁶³. Particularly at the $3d$ cation $L_{3,2}$ edge, orbital polarization extracted from XLD is used as a measure of the oxygen octahedral distortion in perovskites owing to the anisotropic hybridization between cation $3d$ and O $2p$ orbitals⁶⁴. Accordingly, in fluorite-structure ferroelectrics, the magnitude of XLD present at the Zr $M_{3,2}$ edges can be a gauge of the degree of polyhedral distortion (in this case, a distortion of the oxygen tetrahedron) and the oxygen atomic asymmetry. Indeed, the orbital polarization at the Zr M_2 edge is enhanced as the thickness is reduced from the thick (100-cycle) to ultrathin (ten-cycle) regime (Fig. 3c), consistent with diffraction-based results demonstrating amplified structural distortions in the ultrathin limit. Spectroscopy can also help understand the evolution to highly textured films in the ultrathin limit (Fig. 3e, f), as XLD enables both element- and orbital-specific information by comparing polarization-dependent XAS spectra. GI-XRD across the thickness series (Fig. 3f) indicates that the degree of orientation substantially changes as the HZO drops below about 2.5 nm (25 cycles). The microstructure change below 25 cycles also manifests as inverted orbital polarization at the oxygen K edge, particularly at the e manifold (Fig. 3b). Absorption of vertically and horizontally polarized light preferentially probes the polar-distortion-split e levels (the x^2y^2 and $3z^2-r^2$ d orbitals); the reversal of XLD sign indicates these levels are inverted with respect to one another. In perovskites, such a change in orbital polarization is often attributed to different signs of tetragonal distortion (c/a) of the oxygen octahedron⁶⁴. Analogously, here the change in microstructure across 20–25 cycles, namely, the emergence of highly oriented films, could allow confinement strain effects to distort the oxygen tetrahedron more coherently along the polar axis. This synergistic effect could potentially explain the enhanced distortions observed in the ultrathin regime.

Nanospectroscopy. PEEM was performed at the Advanced Light Source beamline 11.0.1. X-rays were incident at 30° off grazing, probing just the first few nanometres of film, spanning the entire ten-cycle (1 nm) HZO thickness. Nanospectroscopy point-by-point scans were

employed to spatially resolve XLD contrast; at each specified energy value in the oxygen *K* edge (520–550 eV) regime, PEEM images were taken for both values of the linear polarization (horizontal, vertical) across a 20- μm field of view ($1,000 \times 1,000$ pixel grid). The exposure to high-flux synchrotron X-rays probably depolarized the ultrathin ferroelectric sample as photoelectrons were removed from the surface, as is observed in ultrathin films of BaTiO_3 and other ferroelectrics; PEEM-XLD images (Extended Data Fig. 9f) illustrate nanoscale domains at the energy range corresponding to the polar *e*-split feature. Data processing to extract XLD contrast involved dividing images of opposite linear polarization, which eliminates topography and work function contrast. Topography and work function artefacts contribute at the pre-edge (about 530 eV), whereas the intrinsic orbital anisotropy contributions manifest only at resonance (about 535 eV); the presence of XLD contrast only at resonance confirms the orbital asymmetry origins of XLD contrast in ultrathin HZO. Furthermore, the highly textured nature of the ultrathin films prevents the XLD contrast from averaging to zero (cancellation would be expected for a fully polycrystalline film) on a length scale smaller than the experimental resolution.

Optical spectroscopy

SHG and inversion asymmetry. Nonlinear optical SHG was performed using a custom setup at University of California, Berkeley, as detailed in a previous work²³. The excitation light was extracted using an optical parametric oscillator (Inspire HF 100, Spectra Physics, Santa Clara) pumped by a mode-locked Ti:sapphire oscillator. The excitation laser was linearly polarized by a 900–1,300 nm polarizing beamsplitter. The transmitted *p*-polarized laser light can change its polarization by rotating an infrared half waveplate before pumping the sample. The laser is focused by a 50 \times near-infrared objective onto the sample. The SHG signal was detected in the backscattering configuration, analysed by a visible-range polarizer, and finally collected by a cooled charge-coupled device spectrometer. SHG was performed with a 960-nm pump and detected at 480 nm under tilt incidence. SHG is commonly used to investigate piezoelectric and ferroelectric single crystals and thin films⁶⁵ as the photon frequency-doubling process is allowed only in materials lacking inversion symmetry.

Field-dependent SHG. Electric-field-dependent SHG experiments were performed on the bare surface of ten-cycle (1 nm) HZO films (top metal was etched away after phase annealing). The HZO layer was then patterned into micrometre-sized islands to enable systematic identification of specific HZO regions; various islands were poled with an electric field (applied by a PFM tip), while other islands were left as is. The optical microscope identified the poled and unpoled islands, and the second harmonic signal was detected across various islands. Increased SHG intensity, sensitive to out-of-plane polarization in this tilt-incidence experimental geometry, in poled HZO islands suggests that the electric field increases the projection of out-of-plane polarization by aligning domains with different polarization directions.

Electrical characterization

Tunnel current measurements. Tunnel current measurements were performed using a commercial Semiconductor Device Analyzer (Agilent B1500) with a pulse generator unit to enable voltage pulses down to the microsecond regime. Samples were patterned into capacitors of various area, with W as the top electrode, and heavily doped Si (10^{19} cm^{-3}) as the bottom contact. The 19- μm W tips (DCP-HTR154-001, FormFactor) made electrical contact within a commercial probe station (Cascade Microtech). In tandem, conducting atomic force microscopy measurements were performed using a commercial scanning probe microscope (Asylum MFP-3D) at University of California, Berkeley. Current–voltage characteristics through the capacitor device were probed in the AFM by using a Keithley 2400 Source Measure Unit to bias the top electrode of the sample through 20-nm-radius Pt/Ir-coated AFM

probes (25PtIr300B cantilever probe, Rocky Mountain Nanotechnology), grounded to the heavily doped Si substrate.

Current–voltage hysteresis and tunnel electroresistance. We used voltage-polarity-dependent current–voltage hysteresis to rule out resistive switching mediated by dielectric breakdown and filamentary-type switching. For filamentary-mediated resistive switching—often observed in amorphous HfO_2 —the sense of hysteresis is dependent on the direction of the voltage sweep (that is, the initial polarity of the voltage waveform), which dictates the filament formation⁶⁶. Meanwhile, ferroelectric tunnel junctions demonstrate the same sense of current–voltage hysteresis independent of the sweep direction; this voltage-polarity independence is indicative of polarization-mediated switching, as observed for our ultrathin ten-cycle (1 nm) HZO films (Extended Data Fig. 10e). To further investigate the origin of the resistive switching, tunnelling electroresistance hysteresis maps as a function of write voltage (at low read voltage) demonstrate saturating, abrupt hysteretic behaviour (Extended Data Fig. 10b, d) characteristic of polarization-driven switching^{67,68}. Evidence of polarization-driven resistive switching from tunnelling electroresistance is provided for ten-cycle (1 nm) HZO films of two different compositions (Extended Data Fig. 10). Although many of the results presented here are for films with 4:1 Hf:Zr ratio, for comparison, we have included results and demonstrated ferroelectricity for a ten-cycle (1 nm) film with this modified 1:1 Hf:Zr ratio. Pioneering work on HZO in the thicker regime ($>5 \text{ nm}$)^{50,69,70} has shown that a 1:1 Hf:Zr ratio often demonstrates the best ferroelectric properties. Ferroelectric tunnel junctions based on composite ferroelectric-dielectric barriers using HZO in this thicker regime demonstrate promising polarization-driven resistive switching results^{71,72}. Optimizing ferroelectric tunnel junction behaviour employing HZO in the ultrathin regime (around 1 nm) will need to be carefully studied.

Data availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

42. Karbasian, G. et al. Stabilization of ferroelectric phase in tungsten capped $\text{Hf}_{0.8}\text{Zr}_{0.2}\text{O}_2$. *Appl. Phys. Lett.* **111**, 022907 (2017).
43. Rodriguez, B. J., Callahan, C., Kalinin, S. V. & Proksch, R. Dual-frequency resonance-tracking atomic force microscopy. *Nanotechnology* **18**, 475504 (2007).
44. Jesse, S., Lee, H. N. & Kalinin, S. V. Quantitative mapping of switching behavior in piezoresponse force microscopy. *Rev. Sci. Instrum.* **77**, 073702 (2006).
45. Hong, S. et al. Principle of ferroelectric domain imaging using atomic force microscope. *J. Appl. Phys.* **89**, 1377–1386 (2001).
46. Strelcov, E. et al. Role of measurement voltage on hysteresis loop shape in piezoresponse force microscopy. *Appl. Phys. Lett.* **101**, 192902 (2012).
47. Leu, C.-C. et al. Domain structure study of $\text{SrBi}_2\text{Ta}_2\text{O}_9$ ferroelectric thin films by scanning capacitance microscopy. *Appl. Phys. Lett.* **82**, 3493–3495 (2003).
48. Chernikova, A. et al. Ultrathin $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2$ ferroelectric films on Si. *ACS Appl. Mater. Interf.* **8**, 7232–7237 (2016).
49. Sang, X., Grimley, E. D., Schenk, T., Schroeder, U. & LeBeau, J. M. On the structural origins of ferroelectricity in HfO_2 thin films. *Appl. Phys. Lett.* **106**, 162905 (2015).
50. Böschke, T. *Crystalline Hafnia and Zirconia Based Dielectrics for Memory Applications* PhD thesis, Hamburg University of Technology, <https://cuvillier.de/en/shop/publications/763-crystalline-hafnia-and-zirconia-based-dielectrics-for-memory-applications> (2010).
51. Zhao, C., Roebben, G., Heyns, M. M. & Van der Biest, O. Crystallisation and tetragonal-monoclinic transformation in ZrO_2 and HfO_2 dielectric thin films. *Key Eng. Mater.* **206–213**, 1285–1288 (2001).
52. Kriegner, D., Wintersberger, E. & Stangl, J. xrayutilities: a versatile tool for reciprocal space conversion of scattering data recorded with linear and area detectors. *J. Appl. Cryst.* **46**, 1162–1170 (2013).
53. Lichtensteiger, C., Triscone, J., Junquera, J. & Ghosez, P. Ferroelectricity and tetragonality in ultrathin PbTiO_3 films. *Phys. Rev. Lett.* **94**, 047603 (2005).
54. Park, M. H. et al. A comprehensive study on the structural evolution of HfO_2 thin films doped with various dopants. *J. Mater. Chem. C* **5**, 4677–4690 (2017).
55. Young, A. T. et al. Variable linear polarization from an x-ray undulator. *J. Synch. Rad.* **9**, 270–274 (2002).
56. Jain, A. et al. The Materials Project: a materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
57. Mathew, K. et al. High-throughput computational X-ray absorption spectroscopy. *Sci. Data* **5**, 180151 (2018).

58. Ong, S. P. et al. Python Materials Genomics (pymatgen): a robust, open-source Python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
59. Cho, D.-Y., Jung, H.-S. & Hwang, C. S. Structural properties and electronic structure of $\text{HfO}_2\text{-ZrO}_2$ composite films. *Phys. Rev. B* **82**, 094104 (2010).
60. Wu, M. et al. Strain and composition dependence of orbital polarization in nickel oxide superlattices. *Phys. Rev. B* **88**, 125124 (2013).
61. de Groot, F. et al. Oxygen 1s X-ray-absorption edges of transition-metal oxides. *Phys. Rev. B* **40**, 5715–5723 (1989).
62. de Groot, F. Multiplet effects in X-ray spectroscopy. *Coord. Chem. Rev.* **249**, 31–63 (2005).
63. Cohen, R. E. Origin of ferroelectricity in perovskite oxides. *Nature* **358**, 136–138 (1992).
64. Pesquera, D. et al. Surface symmetry-breaking and strain effects on orbital occupancy in transition metal perovskite epitaxial films. *Nat. Commun.* **3**, 1189 (2012).
65. Denev, S. A., Lummen, T. T. A., Barnes, E., Kumar, A. & Gopalan, V. Probing ferroelectrics using optical second harmonic generation. *J. Am. Ceram. Soc.* **94**, 2699–2727 (2011).
66. Bersuker, G. & Gilmer, D. Metal oxide resistive random-access memory (RRAM) technology. In *Advances in Non-Volatile Memory and Storage Technology* 288–340 (Elsevier, 2014).
67. Chanthbouala, A. et al. Solid-state memories based on ferroelectric tunnel junctions. *Nat. Nanotechnol.* **7**, 101–104 (2012).
68. Gruverman, A. et al. Tunneling electroresistance effect in ferroelectric tunnel junctions at the nanoscale. *Nano Lett.* **9**, 3539–3543 (2009).
69. Müller, J. et al. Ferroelectricity in simple binary ZrO_2 and HfO_2 . *Nano Lett.* **12**, 4318–4323 (2012).
70. Park, M. H. et al. Surface and grain boundary energy as the key enabler of ferroelectricity in nanoscale hafnia-zirconia: a comparison of model and experiment. *Nanoscale* **9**, 9973–9986 (2017).
71. Fujii, S. et al. First demonstration and performance improvement of ferroelectric HfO_2 -based resistive switch with low operation current and intrinsic diode property. In *2016 IEEE Symposium on VLSI Technology* 1–2 (IEEE, 2016).
72. Max, B., Hoffmann, M., Slesazek, S. & Mikolajick, T. Ferroelectric tunnel junctions based on ferroelectric-dielectric $\text{Hf}_{0.5}\text{Zr}_{0.5}\text{O}_2/\text{Al}_2\text{O}_3$ capacitor stacks. In *2018 48th European Solid-State Device Research Conference (ESSDERC)* 142–145 (IEEE, 2018).

Acknowledgements This research was supported in part by the Berkeley Center for Negative Capacitance Transistors (BCNCT), ASCENT (Applications and Systems-Driven Center for Energy-Efficient Integrated NanoTechnologies), one of the six centres in the JUMP initiative

(Joint University Microelectronics Program), an SRC (Semiconductor Research Corporation) programme sponsored by DARPA, the DARPA T-MUSIC (Technologies for Mixed-mode Ultra Scaled Integrated Circuits) programme and the UC MRPI (University of California Multicampus Research Programs and Initiatives) project. This research used resources of the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under contract number DE-AC02-06CH11357. This research used resources of the Advanced Light Source, which is a DOE Office of Science User Facility under contract number DE-AC02-05CH11231. Use of the Stanford Synchrotron Radiation Light source, SLAC National Accelerator Laboratory, is supported by the US DOE, Office of Science, Office of Basic Energy Sciences under contract number DE-AC02-76SF00515. Electron microscopy was performed at the Molecular Foundry, LBNL, supported by the Office of Science, Office of Basic Energy Sciences, US DOE (DE-AC02-05CH11231). J.C. and R.d.R. acknowledge additional support from the Presidential Early Career Award for Scientists and Engineers (PECASE) through the US DOE. J.X. and X.Z. acknowledge support from the National Science Foundation (NSF) under grant 1753380 and the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research award OSR-2016-CRG5-2996.

Author contributions Film synthesis was performed by S.S.C., G.K. and D.K. Device fabrication was performed by D.K. Electron microscopy was performed by R.d.R. and S.-L.H. under the supervision of J.C. and R.R., respectively, and analysis was performed by L.-C.W. under the supervision of S.S. Scanning probe microscopy was performed by S.S.C. and N.S. IDS measurements were performed and developed by R.W. and R.P. SCM was performed by H.Z. X-ray structural characterization was performed by S.S.C., N.S. and M.R.M. under the supervision of A.M. and E.K. X-ray spectroscopy and microscopy was performed by S.S.C. under the supervision of R.V.C., P.S. and E.A. Second harmonic generation was performed by J.X. under the supervision of X.Z. Electrical measurements were performed by S.S.C., N.S. and A.D. S.S.C. and S.S. co-wrote the manuscript. S.S. supervised the research. All authors contributed to discussions and commented on the manuscript.

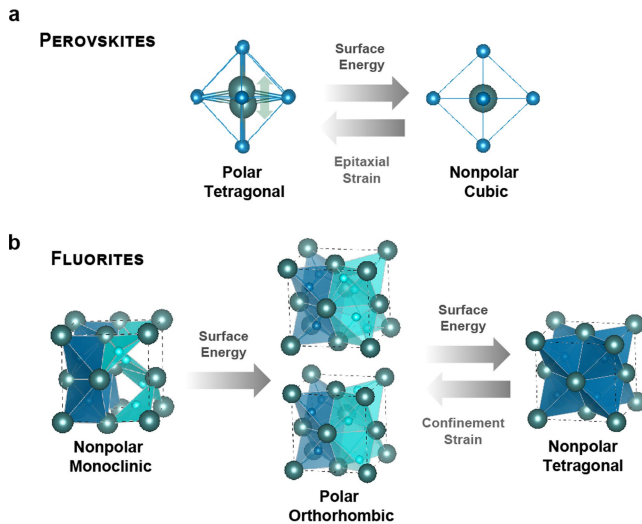
Competing interests The authors declare no competing interests.

Additional information

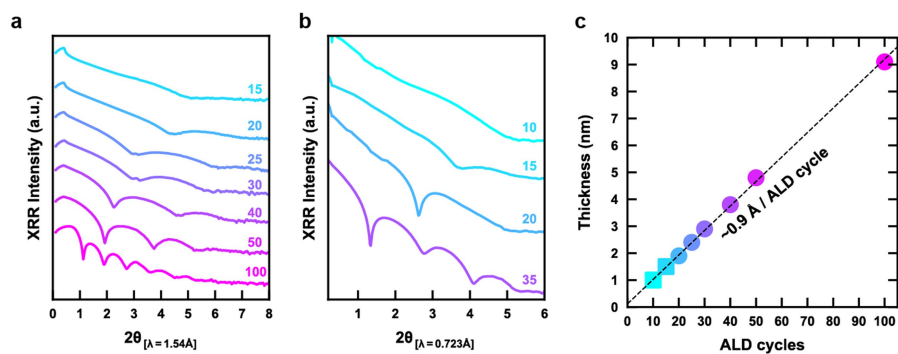
Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2208-x>.

Correspondence and requests for materials should be addressed to S.S.C. or S.S.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

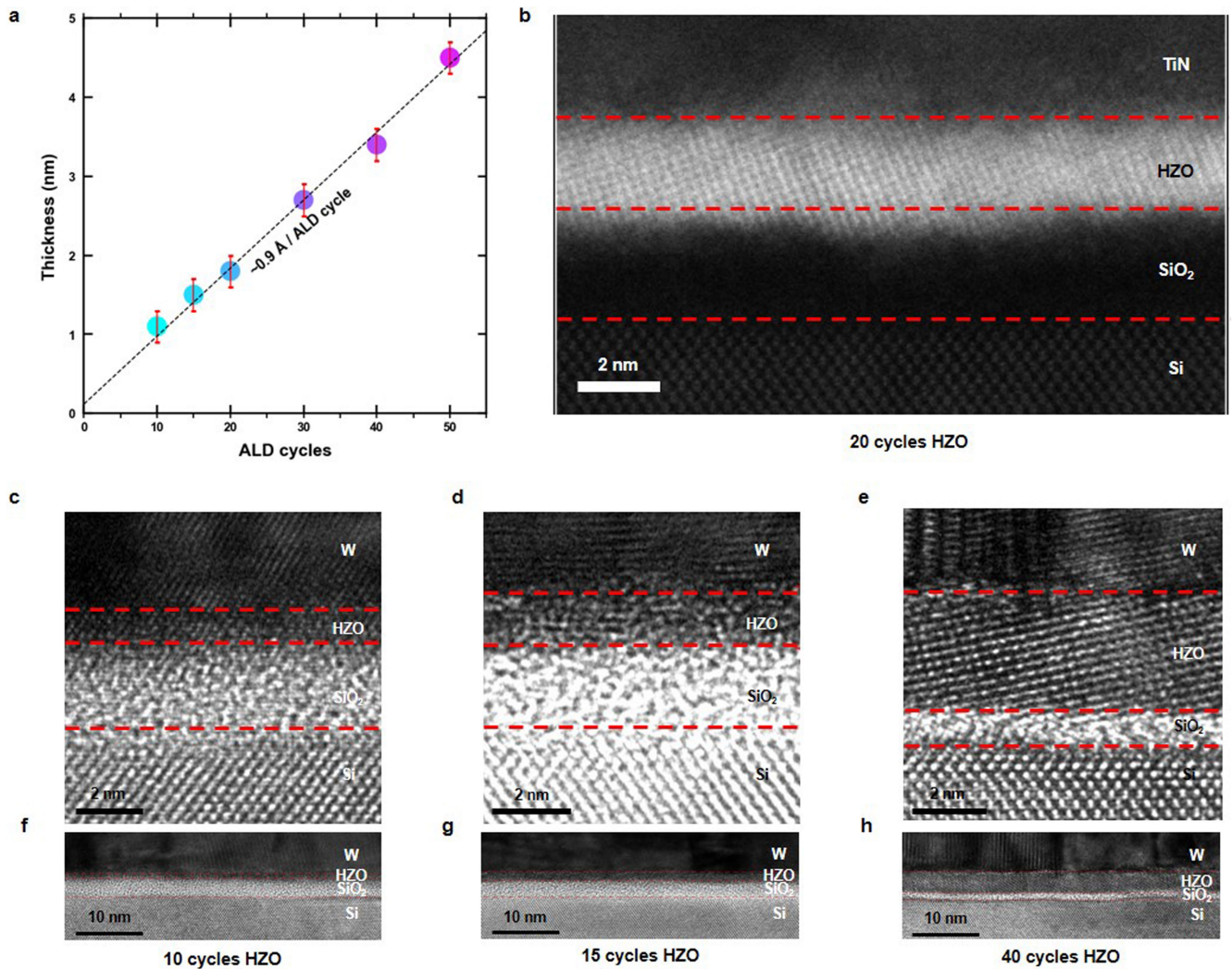


Extended Data Fig. 1 | Size effects in fluorite- and perovskite-structure ferroelectrics. a, In perovskite ferroelectrics, the polar ‘tetragonal’ distortion (c/a) can be represented as the centre cation displacement with respect to its surrounding oxygen octahedron. **b,** In fluorite-structure ferroelectrics, the polar ‘rhombic’ distortion ($2c/(a+b)$) can be represented as the centre anion displacement with respect to its surrounding cation tetrahedron; in the nonpolar T-phase, the oxygen atom (blue) lies in the polyhedral centre of the tetrahedron. The evolution of the bulk-stable M-phase to the high-symmetry T-phase and polar O-phase in the fluorite-structure structure illustrates the role of size effects (surface energies favour higher symmetry) and confinement strain (distortions favour lower symmetry) on stabilizing inversion asymmetry. Surface energies are critical when considering the role of size effects on ferroelectricity; higher-symmetry phases are energetically favourable at reduced dimensions owing to lower unit cell volumes. In fluorite structures (perovskites), the noncentrosymmetric O-phase (T-phase) has higher (lower) symmetry than the bulk-stable centrosymmetric M-phase (C-phase). Consequently, surface energies help to counteract depolarization fields in fluorite-structure ferroelectrics—already diminished in fluorite structures relative to perovskites owing to its lower dielectric constant⁴—in the ultrathin regime. Therefore, both intrinsic (surface energies) and extrinsic (confinement strain) mechanisms can favour ultrathin inversion symmetry breaking in fluorite structures. Meanwhile, both surface and depolarization energies tend to destabilize inversion asymmetry in perovskite ferroelectrics, while epitaxial strain can stabilize symmetry-lowering polar distortions³⁶.



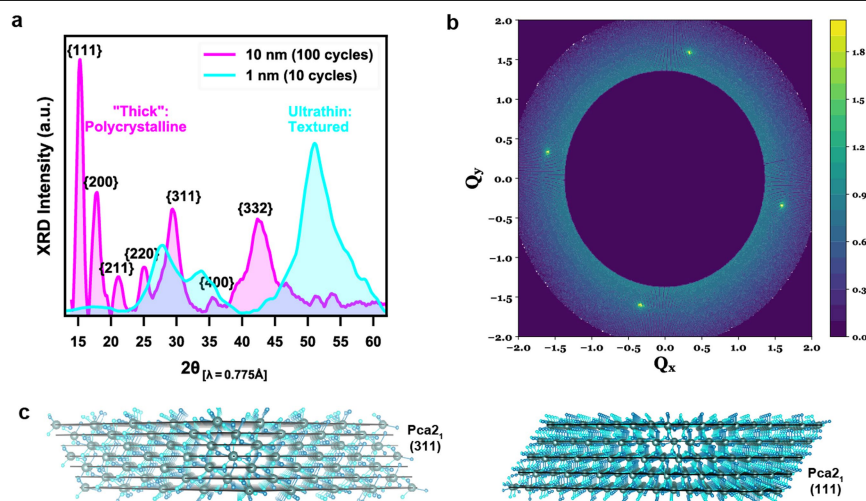
Extended Data Fig. 2 | Thickness verification of ultrathin HZO films from XRR. **a**, Laboratory diffractometer XRR of HZO thickness series, demonstrating clear fringes for thickness extraction present down to 20-cycle HZO. **b**, Synchrotron XRR of ultrathin HZO films, enabling thickness fitting analysis for sub-20-cycle films. **c**, HZO thickness as a function of ALD cycles, as

determined by fitting analysis from XRR. The growth rate is about 11 cycles nm^{-1} , verified across 10–100 ALD cycle films. Squares (circles) represent thicknesses extracted from fitting to synchrotron (laboratory diffractometer) XRR measurements.



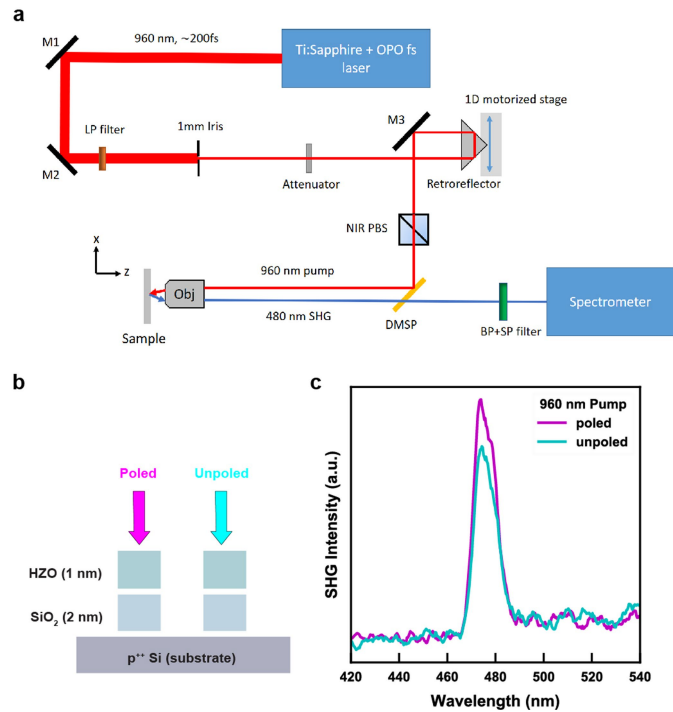
Extended Data Fig. 3 | Thickness verification of ultrathin HZO films using TEM. **a**, HZO thickness as a function of ALD cycles, as determined by Si atomic lattice calibration from TEM imaging. The growth rate is $\sim 11 \text{ cycles nm}^{-1}$, verified across 10–50 ALD cycle films, consistent with XRR (Extended Data Fig. 2). The red error bars reflect 2σ variation. **b**, Cross-sectional ADF STEM

image of 20 cycles HZO. **c–e**, Cross-section TEM images of ten-cycle HZO (**c**), 15-cycle HZO (**d**) and 40-cycle HZO (**e**). **f–h**, Wide field-of-view TEM images of ten-cycle HZO (**f**), 15-cycle HZO (**g**) and 40-cycle HZO (**h**) to provide a perspective of the heterostructure uniformity. The Si substrate is oriented along the $[110]$ zone axis for all TEM images.

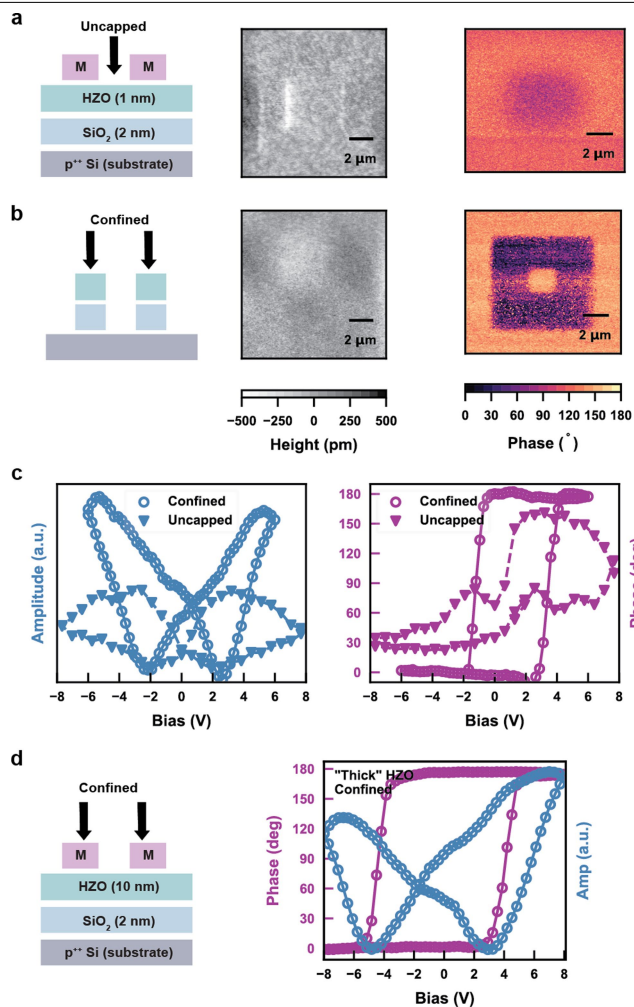


Extended Data Fig. 4 | Emergence of highly-textured films in the ultrathin regime. **a**, Synchrotron GI-XRD scans ($\lambda = 0.775 \text{ \AA}$) of HZO thickness series endmembers: 10-cycle and 100-cycle. The 100-cycle HZO film is indexed according to the polar orthorhombic phase $Pca2_1$. Many of the polycrystalline reflections, most notably the (111), are no longer present at an appreciable intensity in the ultrathin limit owing to the geometric constraints of one-dimensional spectra (unable to probe all reflections present in highly oriented films) (Methods). Instead tilted-geometry diffraction (pole figures) are used to access the oriented reflections. **b**, Pole figure of ten-cycle HZO, taken at a Q_x slice corresponding to the film (111) lattice spacing. The radial direction represents χ , while the azimuthal direction represents ϕ (0° – 360° range). The presence of four intense peaks corresponding to the four

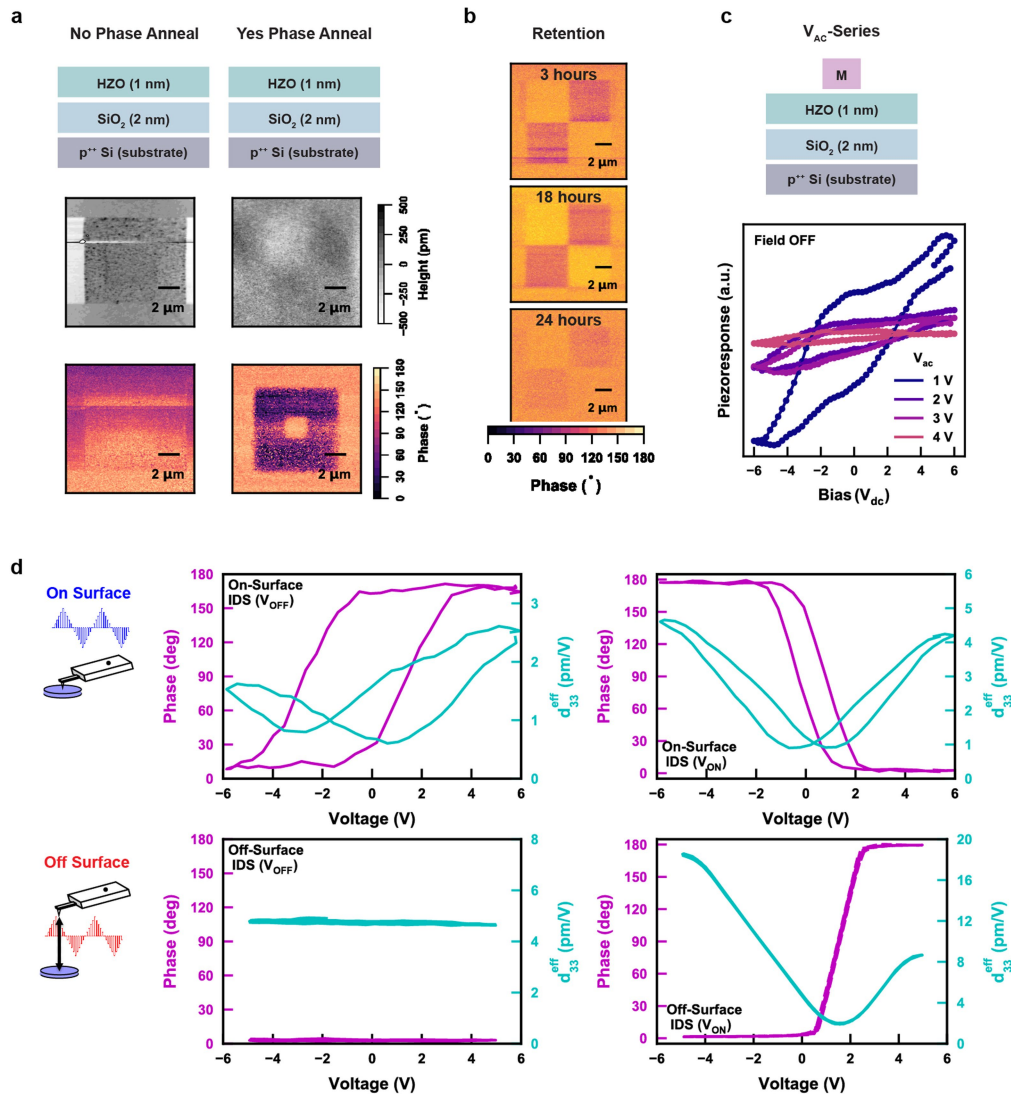
(111)-projections indicate the highly textured nature of the ultrathin HZO film. The four Si (111)-projections would be expected at $\phi = 45^\circ$ off from the $Q_{x,y}$ principal axes at a smaller value of Q_x . **c**, Schematic of the (311) (left) and (111) (right) close-packed planes in the fluorite-structure structure. All the cation sites lie on such planes, which minimize surface energy effects because only metal-oxygen dangling bonds are present out-of-plane. We note that all schematics reflect stacking of the respective planes to a total thickness of 1 nm, although ultrathin HZO films may not exhibit such stacking throughout the film. For ten-cycle films, {311} indexing is consistent with the relevant intensity (about 30°) observed in the out-of-plane one-dimensional GI-XRD pattern (**a**), and the (111) reflections are present from the two-dimensional pole figure pattern (**b**).



Extended Data Fig. 5 | Inversion symmetry breaking in ultrathin HZO via SHG. **a**, Schematic of the SHG experimental setup, using a 960-nm pump and SHG intensity detected around 480 nm under tilt incidence, which is sensitive to out-of-plane polarization (Methods). NIR, near-infrared; 1D, one-dimensional; PBS, polarized beam splitter; Obj, objective; LP, BP and SP represent long-pass, band-pass and short-pass filters; DMSP, dichroic short-pass mirror; M1, M2 and M3 refer to mirrors; OPO, optical parametric oscillator. **b**, Schematic of the ten-cycle HZO islands probed by SHG (Methods); micrometre-sized islands enabled identification of specific HZO regions either poled with an electric field (applied by a PFM tip) or left as is. For these experiments, heavily doped (10^{19} cm^{-3}) p-type Si substrates (p⁺⁺ Si) are used to serve as the bottom electrode. **c**, SHG spectrum on a ten-cycle HZO film, comparing poled versus unpoled SHG intensity. Spontaneous polarization is demonstrated by the presence of SHG—allowed only for inversion asymmetric systems—in unpoled ten-cycle HZO. This is consistent with PFM phase contrast in unpoled HZO regions (Fig. 2c), indicating elimination of the ‘wake-up’ effects for ferroelectricity in ultrathin HZO. The enhanced SHG contrast in poled films—possibly due to the electric field converting a small fraction of the film to the polar phase or aligning polar domains—indicates that the mechanism behind the SHG contrast is field-tunable. This field-enhanced SHG is consistent with ferroelectric origins and would probably eliminate SHG contrast from surface effects.

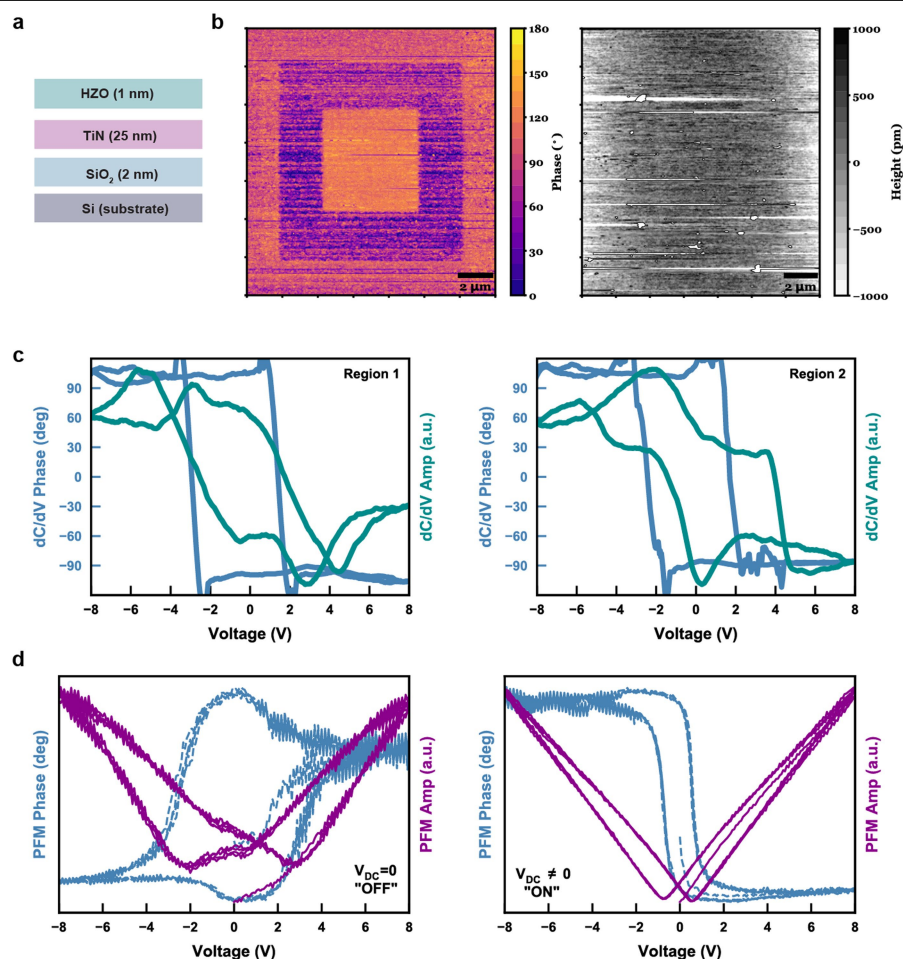


Extended Data Fig. 6 | Role of ultrathin confinement for polar phase stabilization. **a, b**, Schematic structure (left) probed by PFM (tip location indicated by arrows), topography (centre), and PFM phase contrast images (right) on ten-cycle HZO in a region that was uncapped (**a**) versus confined (**b**) by W (represented by 'M' for metal in the schematic) during phase annealing. Robust 180° phase contrast is only present for the confined HZO. **c**, Phase (left) and amplitude (right) switching spectroscopy loops ($V_{dc} = 0$, 'OFF' state) as a function of bias voltage on ten-cycle HZO films, demonstrating the critical role of confinement during phase annealing in stabilizing ferroelectricity in ultrathin HZO. 180° phase contrast and butterfly-shaped amplitude are present only for confined HZO. Therefore, both switching-spectroscopy PFM and PFM imaging illustrate the critical role of confinement during phase annealing for stabilizing the ferroelectric phase. For the PFM images, ± 7 V was applied in a 'box-in-box' poling pattern directly on the HZO surface, and switching-spectroscopy PFM loops were measured on capacitor structures (Methods). **d**, Schematic structure (left) probed by PFM (tip location indicated by arrows) and PFM phase and amplitude hysteresis loops (right) as a function of bias voltage on 100-cycle HZO in a region that was confined by W during phase annealing. Thicker 100-cycle HZO also demonstrates ferroelectric behaviour.



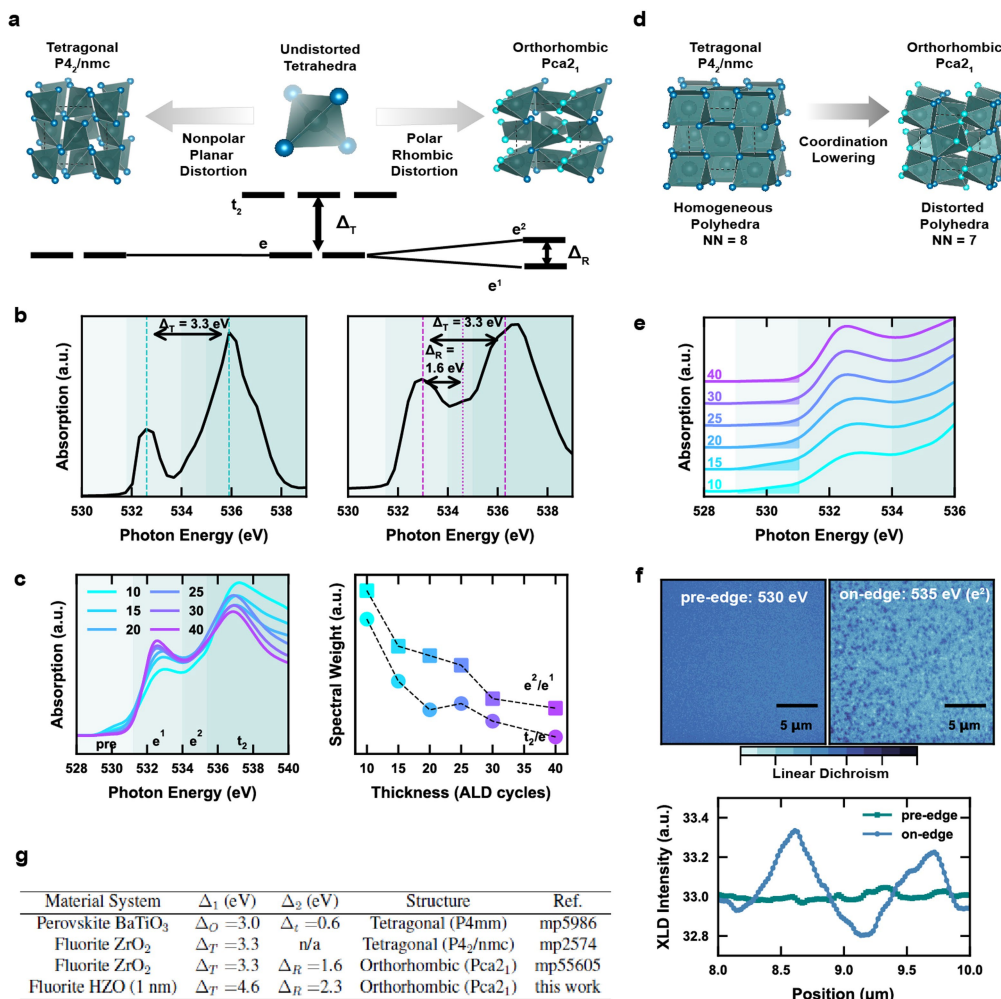
Extended Data Fig. 7 | Eliminating artefacts from scanning probe microscopy. **a**, Topography and PFM phase contrast images for ten-cycle HZO which did not (left) and did (right) undergo annealing after ALD deposition. The terraced topography in the non-annealed film indicates that the weak phase contrast is falsely caused by field-induced topographic changes. This is consistent with charge injection or ion migration, which plague amorphous HfO₂ films²⁵. Phase-annealed films do not display such field-induced topographic distortions yet demonstrate much clearer phase contrast, indicating the origin of PFM phase contrast in crystalline HZO films is different than that of amorphous HZO films. In the images shown, ± 7 V were applied in a ‘box-in-box’ poling sequence. **b**, Time-dependent PFM phase contrast images on a ten-cycle HZO film across a 24-h period. In the images shown, ± 7 V was applied in the indicated checkerboard poling pattern. **c**, Collapse of the PFM loop from V_{ac}-series. Schematic capacitor structure probed by PFM (top) and piezoresponse as a function of V_{ac} in the ‘OFF’ (V_{dc} = 0) state (bottom),

demonstrating the collapse of the PFM loop as V_{ac} approaches the coercive voltage. This provides further confirmation of the ferroelectric origin of the PFM signal as opposed to tip bias-induced mechanisms⁴⁶. The non-ideal shape of the piezoresponse loops, particularly at higher voltages, is probably caused by non-ferroelectric contributions from the additional dielectric SiO₂ layer through which most of the voltage is dropped (Methods). **d**, IDS switching-spectroscopy measurements on ten-cycle (1 nm) HZO, demonstrating hysteresis for the PFM tip on-surface (top) versus no hysteresis for the tip off-surface (bottom). The on-surface loops indicate 180° phase hysteresis and butterfly-shaped d_{33}^{eff} , indicative of ferroelectric behaviour. IDS PFM measurements (Methods) remove the long-range electrostatics and cantilever resonance artefacts that plague typical voltage-modulated PFM switching spectroscopy²⁶. This ferroelectric origin of the hysteresis is further supported by non-hysteretic off-surface loops²⁶, which probe electrostatic contributions.



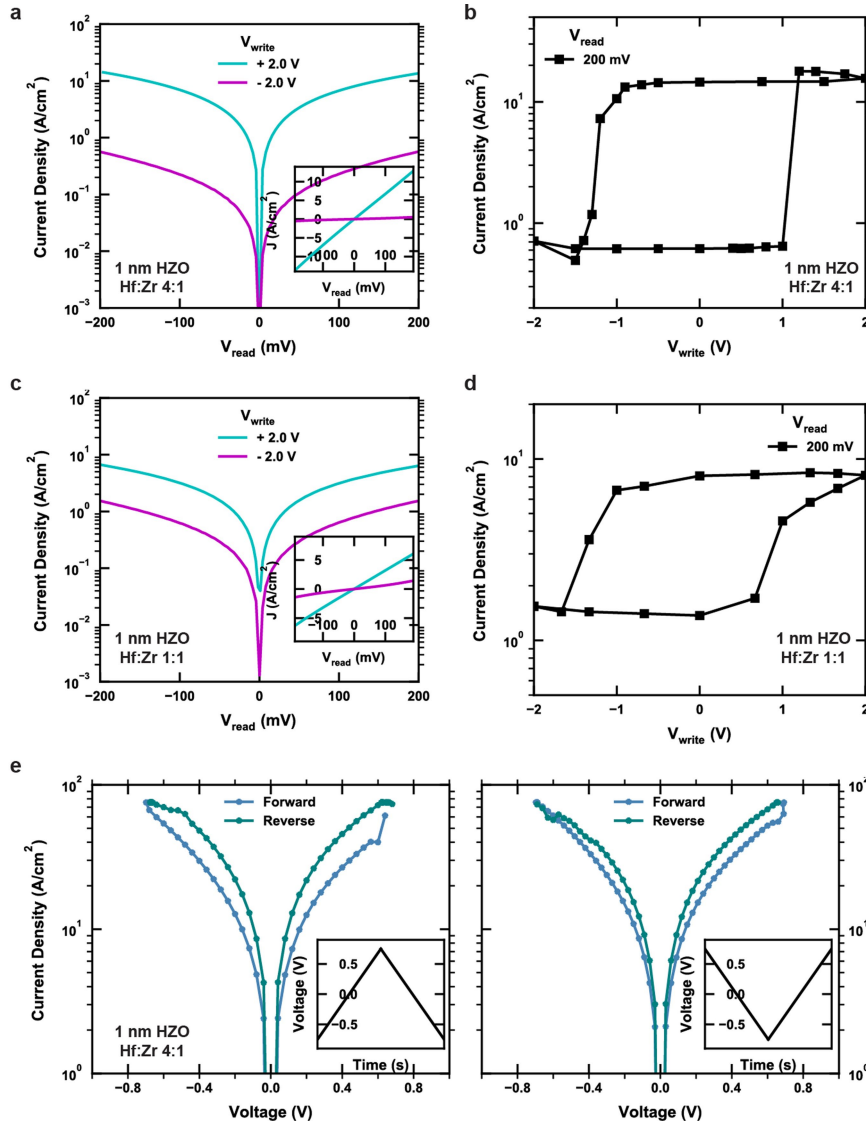
Extended Data Fig. 8 | High-frequency capacitance characterization of ultrathin HZO. **a**, Schematic heterostructure of ultrathin HZO on metallic TiN probed the microwave capacitance measurements to eliminate contributions from the semiconducting Si substrate. **b**, PFM phase contrast (left) and topography (right) imaging for 10 cycles HZO on TiN-buffered Si. Ultrathin ferroelectricity persists on top of metallic underlayers as well as dielectric SiO₂, although the topography is rougher than the films on SiO₂ due to the inhomogeneity introduced by the sputtered TiN. **c**, SCM dC–dV spectroscopy loops taken on multiple bare regions of an ultrathin ten-cycle HZO film, demonstrating reproducible SCM response. The square 180° phase hysteresis

and dC/dV loops, which integrates into the classic butterfly-shaped capacitance–voltage plot (Fig. 2c), provides conclusive evidence of ferroelectric polarization switching beyond PFM loops (Fig. 2e, Extended Data Fig. 6). The microwave-frequency nature of the SCM enables leakage-mitigated differential capacitance measurements of ultrathin films (Methods). **d**, PFM switching-spectroscopy loops taken on the same region of the ten-cycle HZO as the SCM measurements, confirming the ferroelectric-like phase and amplitude hysteresis. We note that the SCM and PFM switching spectroscopy was done using the Asylum Cypher scanning probe microscope at Asylum Research (Methods).



Extended Data Fig. 9 | Ultrathin-enhanced distortions and polar signatures from spectroscopy. a, Crystal field splitting diagram for the fluorite-structure structural polymorphs; symmetry-induced e -splitting provides a spectroscopic signature for the polar O-phase (Methods). **b**, Delineating symmetry-split energy regimes in oxygen K -edge XAS. Just as convergent beam electron diffraction provides signatures to demonstrate inversion symmetry breaking⁴⁹, XAS provides spectroscopic signatures to distinguish between the nonpolar tetragonal and polar orthorhombic polymorphs (difficult to resolve from GI-XRD). Left, simulated XAS spectrum for tetragonal ZrO₂ ($P4_2/nmc$) and right, polar orthorhombic ZrO₂ ($Pca2_1$), both courtesy of the Materials Project^{56,57}. The background colour shading denotes the symmetry-split regimes explained in the crystal field splitting diagram. **c**, Experimental XAS data on ultrathin HZO displays similar spectroscopic XAS features as the simulated polar O-phase ($Pca2_1$)—namely, relative e/t_2 spectral weight and splittings corresponding to tetrahedral (Δ_T) and rhombic (Δ_R) distortions. Left, XAS of the HZO thickness series at the O K -edge, zooming in on the e - and t_2 -regimes. Right, O K -edge spectral weight trends as a function of HZO thickness. The relative spectral weights from the t_2/e and e -split regimes indicate enhanced tetrahedral (Δ_T) and rhombic distortions (Δ_R) in ultrathin films, respectively, consistent with C_{2v} symmetry of the polar O-phase. **d**, Schematic representation of the cation nearest-neighbour coordination dropping from NN = 8 (T-phase) to NN = 7 (polar O-phase) as the crystal symmetry is lowered. The disorder in oxygen polyhedral coordination (note

the different oxygen atoms denoted by the blue and cyan atoms in the polar O-phase) manifests as spectral weight in the pre-edge regime⁶². **e**, The experimental pre-edge spectral weight as a function of thickness, indicating ultrathin-enhanced polyhedral disorder. **f**, Top: PEEM-XLD images of ten-cycle (1 nm) HZO at the O K -edge. Pre-edge images (left) exhibit no XLD contrast, while on-edge images (right)—at the energy corresponding to the polar-distortion split e -regime—demonstrate XLD contrast. This suggests that XLD is indeed sensitive to polar features in ultrathin highly textured HZO. Bottom, line profile of the XLD intensity, demonstrating substantial variations in on-edge XLD data compared to noise for pre-edge XLD. **g**, Crystal field splitting energies in HZO-related transition metal oxide systems. The material system, primary crystal electric field (Δ_1), secondary crystal electric field (Δ_2), and structure for various systems related to HZO and perovskite ferroelectrics are shown, where Δ_O , Δ_t , Δ_T and Δ_R corresponds to octahedral, tetragonal, tetrahedral, and rhombic crystal electric field (CEF), respectively. The reference crystal electric field values are taken from the Materials Project database⁵⁷ (reference codes denoted by ‘mp’), and the experimental values are extracted via XAS energy-split features (**b**). The large tetrahedral (Δ_T) and rhombic (Δ_R) crystal field splitting energies present in ten-cycle HZO films are much larger than expected values for the polar fluorite-structure ZrO₂ (**b**), which highlights the enhanced distortion present in ultrathin films subject to confinement strain, and is consistent with anomalously large structural distortions extracted from diffraction (Fig. 3g).



Extended Data Fig. 10 | Ultrathin HZO ferroelectric tunnel junction. **a, c,** Tunnel current-voltage characterization of $\text{Si}(\text{p}^{++})/\text{SiO}_2(1\text{ nm})/\text{HZO}(\sim 1\text{ nm})/\text{W}$ capacitor devices—demonstrated for ten-cycle HZO with Hf:Zr composition 4:1 (**a**) and 1:1 (**c**)—as a function of the write pulse (to set the ferroelectric polarization state). Tunneling electroresistance behaviour is demonstrated for $\pm 2\text{ V}$ write and 100 mV read. Insets, linear-scale current-voltage characteristics of the two polarization-driven current states. **b, d,** Tunneling electroresistance hysteresis map as a function of write voltage (demonstrated for ten-cycle HZO with Hf:Zr composition 4:1 (**b**) and 1:1 (**d**)) measured at 200 mV read voltage. The abrupt hysteretic behaviour and saturating tunnelling electroresistance is characteristic of polarization-driven switching⁶⁷, as

opposed to filamentary-based switching caused by electrochemical migration and/or oxygen vacancy motion (Methods). **e,** Current-voltage hysteresis sweeps ruling out non-polarization-driven resistive switching mechanisms (Methods). The device demonstrates current-voltage hysteresis at low voltage and voltage polarity-independent current-voltage hysteresis sense: both negative-positive-negative voltage polarity (left) and positive-negative-positive voltage polarity (right) demonstrate counter-clockwise hysteresis. Such behaviour rules out resistive switching mediated by dielectric breakdown and filamentary mechanisms⁶⁶ and is consistent with polarization-driven switching.

Non-volatile electric control of spin–charge conversion in a SrTiO₃ Rashba system

<https://doi.org/10.1038/s41586-020-2197-9>

Received: 20 August 2019

Accepted: 25 February 2020

Published online: 22 April 2020

 Check for updates

Paul Noël^{1,4,6}, Felix Trier^{2,6}, Luis M. Vicente Arche², Julien Bréhin², Diogo C. Vaz^{2,5}, Vincent Garcia², Stéphane Fusil^{2,3}, Agnès Barthélémy², Laurent Vila¹, Manuel Bibes^{2,3} & Jean-Philippe Attané¹✉

After 50 years of development, the technology of today's electronics is approaching its physical limits, with feature sizes smaller than 10 nanometres. It is also becoming clear that the ever-increasing power consumption of information and communication systems¹ needs to be contained. These two factors require the introduction of non-traditional materials and state variables. As recently highlighted², the remanence associated with collective switching in ferroic systems is an appealing way to reduce power consumption. A promising approach is spintronics, which relies on ferromagnets to provide non-volatility and to generate and detect spin currents³. However, magnetization reversal by spin transfer torques⁴ is a power-consuming process. This is driving research on multiferroics to achieve low-power electric-field control of magnetization⁵, but practical materials are scarce and magnetoelectric switching remains difficult to control. Here we demonstrate an alternative strategy to achieve low-power spin detection, in a non-magnetic system. We harness the electric-field-induced ferroelectric-like state of strontium titanate (SrTiO₃)^{6–9} to manipulate the spin–orbit properties¹⁰ of a two-dimensional electron gas¹¹, and efficiently convert spin currents into positive or negative charge currents, depending on the polarization direction. This non-volatile effect opens the way to the electric-field control of spin currents and to ultralow-power spintronics, in which non-volatility would be provided by ferroelectricity rather than by ferromagnetism.

'Spin–orbitronics'¹² exploits the interplay between charge currents and spin currents enabled by spin–orbit coupling (SOC) in non-magnetic systems. It allows the generation of pure spin currents from charge currents and vice versa, without resorting to ferromagnetic materials. The Edelstein effect¹³ allows charge–spin conversion¹⁴ with an efficiency comparable to or larger than that of the spin Hall effect¹⁵. It typically occurs at Rashba surfaces and interfaces¹⁶ where inversion symmetry breaking results in an out-of-plane electric field. In the presence of SOC, this leads to a locking of the momentum and spin degrees of freedom. The flow of an in-plane charge current in such a system produces a transverse spin density, which can diffuse as a spin current in an adjacent material¹³. Conversely, injecting a spin density results in the production of a net charge current by the inverse Edelstein effect¹⁷. As such, Rashba systems can be used as spin generators and detectors. However, the conversion rate is inherently set by the electronic structure, and cannot be switched by an external stimulus.

The order parameter of ferroelectrics (polarization) can be switched by an electric field for energy costs typically 1,000 times smaller² than those for switching ferromagnets. Moreover, ferroelectrics can harbour intense electric fields, substantially modifying the carrier densities in adjacent materials, and thereby tuning their properties in a non-volatile fashion. An exciting route towards low-power electronics would thus be

to combine the remanence of ferroelectrics with the ability to generate and manipulate spin currents by the direct and inverse Edelstein effects in Rashba systems. Beyond magnetoelectricity, ferroelectric Rashba architectures would therefore offer a new approach to the non-volatile control of spin currents by electric fields, with ultralow-power operation.

Most efforts to identify single-phase Rashba ferroelectrics¹⁸ have focused on GeTe (ref. ¹⁹). However, ferroelectric properties are poor²⁰ because of high leakage, and spin–charge conversion experiments have yielded a moderate efficiency²¹. Here we show that beyond bulk materials, interface systems combining Rashba SOC and a switchable polarization enable the non-volatile electric control of a highly efficient spin–charge conversion.

The general concept of ferroelectricity-controlled spin–charge conversion is described in Fig. 1. At the interface between a ferroelectric and an ultrathin SOC system—such as a heavy metal, a Weyl semi-metal, or a two-dimensional electron gas (2DEG)—electrons are accumulated or depleted depending on the polarization direction (Fig. 1a). This modifies the electric field in the interface region, and in the ideal case changes its sign. If a Rashba state is present in the SOC system at the interface with the ferroelectric, reversing the sign of the local electric field reverses the chirality of the spin textures in both split

¹Université Grenoble Alpes, CEA, CNRS, Spintec, Grenoble, France. ²Unité Mixte de Physique, CNRS, Thales, Université Paris-Saclay, Palaiseau, France. ³Université d'Evry, Université Paris-Saclay, Evry, France. ⁴Present address: ETH Zürich, Zurich, Switzerland. ⁵Present address: CIC Nanogune, Donostia–San Sebastian, Spain. ⁶These authors contributed equally: Paul Noël, Felix Trier.

✉e-mail: manuel.bibes@cnrs-thales.fr; jean-philippe.attane@cea.fr

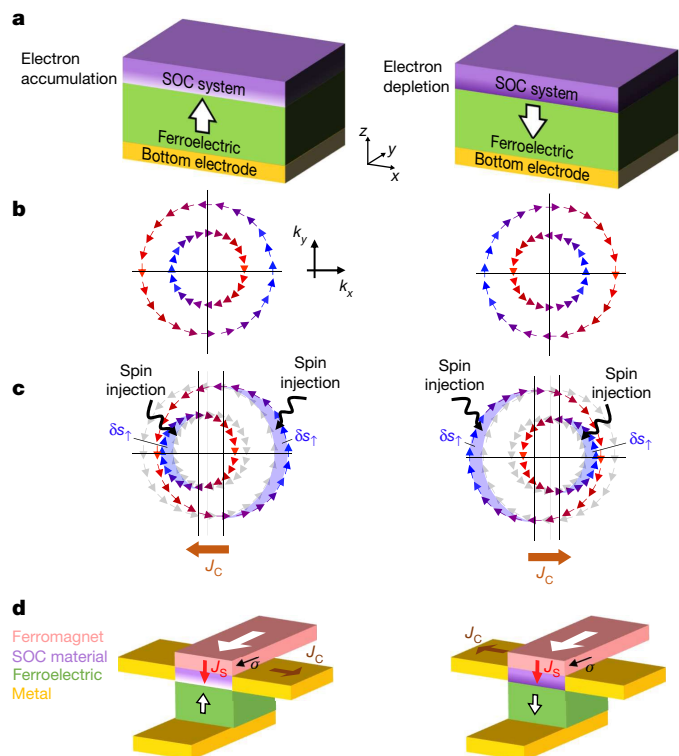


Fig. 1 | Concept of ferroelectricity-controlled spin-charge conversion.

a, Sketch of a ferroelectric Rashba architecture combining a ferroelectric material (green) and a material with spin-orbit coupling (SOC; purple). Upon switching polarization, electrons are accumulated (left) or depleted (right) in the SOC material (for example, a 2DEG), creating an electric field whose sign depends on the polarization direction. **b**, Corresponding Rashba-split chiral Fermi contours with spin-momentum locking. The chirality of the contours switches upon switching the ferroelectric polarization. k_x and k_y correspond to the x and y axes in momentum space. Blue and red colours indicate the spin direction (blue, along k_y ; red, along $-k_y$). **c**, Inverse Edelstein effect in a Rashba interface. When a spin current is injected (for example, by spin pumping) with a spin polarization along the y axis, the spin population is altered, causing a displacement of the two inequivalent Fermi surfaces (red and blue lines) by $\pm \Delta k$ in momentum space. This results in a net charge current generated perpendicularly to the spin current and to its spin polarization. The sign of the generated current depends on the chirality of the Fermi contours and is thus reversed upon switching ferroelectric polarization. δs_\uparrow corresponds to the injected excess of spin-up density. **d**, Non-volatile device operated by ferroelectricity and Rashba SOC. Through the inverse Edelstein effect a charge current J_C is generated by the conversion of a spin current J_S injected from the ferromagnet. The sign of J_C changes with the direction of the ferroelectric polarization. The large white arrows show the ferromagnet magnetization, the small black arrows the spin σ , the brown arrows the direction of the charge current J_C , and the red arrows the direction of the spin current J_S . The small black-and-white arrows correspond to the ferroelectric polarization.

Fermi contours (Fig. 1b). Through the inverse Edelstein effect¹³, the injection of a spin current into the Rashba state will produce a charge current J_C whose sign will depend on the ferroelectric polarization state (Fig. 1c). This mechanism offers the possibility to design a wealth of devices such as the bipolar memory proposed in Fig. 1. It can also be the basis of logic devices²² akin to the magnetoelectric spin-orbit (MESO) device proposed by Intel²³, but without resorting to a multiferroic to switch the ferromagnet.

To experimentally demonstrate the non-volatile electric control of the spin-charge conversion, we use SrTiO₃ (STO) 2DEGs, generated by the deposition of a film of Al onto a STO single crystal^{24,25}. Indeed, STO 2DEGs exhibit a sizeable Rashba SOC¹⁰ with a very high conversion efficiency^{25,26}. In addition, STO is a quantum paraelectric that

develops an electric-field-induced switchable polarization at low temperature⁷⁻⁹.

The spin-to-charge conversion was measured by using spin pumping by ferromagnetic resonance on a NiFe(20 nm)/Al(0.9 nm)/STO sample (see Fig. 2a inset). The nominally 500- μm -thick STO substrate was thinned down to $250 \pm 20 \mu\text{m}$ using mechanical polishing, allowing the application of high electric fields (E). A static magnetic field was applied along the y direction. At the ferromagnetic resonance, a pure spin current is injected into the 2DEG along the $-z$ direction, with spins oriented along y (ref. 27). The measurement of the extra damping due to this relaxation channel allows calculation of the injected spin current^{26,27}. In the 2DEG, this spin current is then converted into a charge current oriented along x by the inverse Edelstein effect. Since the sample is in open circuit, at the resonance field this results in a voltage drop along the sample, in the x direction²⁶.

In the pristine, ungated state, the voltage drop obtained at resonance corresponds to the production of a positive normalized current of $1.2 \text{ A mT}^{-2} \text{ m}^{-1}$ (top left panel of Fig. 2b). At low temperature, STO is known to undergo a phase transition at high electric field⁷⁻⁹: once a large electric field has been applied, the material develops a switchable, remanent polarization. This phenomenon is often referred to as a field-induced ferroelectric order or a field-induced ferroelectric-like state. We applied voltages up to $\pm 200 \text{ V}$, corresponding to E up to $\pm 8 \text{ kV cm}^{-1}$, high enough to achieve this phase transition^{7,9}. After a first initialization cycle [$+200 \text{ V}$; -200 V ; $+200 \text{ V}$], the gate-voltage dependence of the spin-pumping signal shows a hysteretic behaviour (Fig. 2a). The charge currents produced at ferromagnetic resonance have opposite signs for $+200 \text{ V}$ and -200 V gate voltages, as seen in points B, F and D of Fig. 2a and b. After applying the maximum voltage, the normalized current reaches a very high amplitude ($\pm 8.8 \text{ A mT}^{-2} \text{ m}^{-1}$), beyond the record values obtained previously in LaAlO₃/STO and Al/STO samples (around $5 \text{ A mT}^{-2} \text{ m}^{-1}$)²⁶. The spin-charge conversion efficiency is quantified by the inverse Edelstein length λ_{IEE} , equal to the ratio of the 2D charge current density produced by the injected 3D spin current, that is, $\lambda_{\text{IEE}} = J_C^{\text{2D}} / J_S^{\text{3D}}$ (ref. 17; see Methods). Here we estimate $\lambda_{\text{IEE}} \approx 60 \text{ nm}$, a value one to two orders of magnitude larger than in metallic Rashba interfaces¹⁷ or topological insulators²⁸.

We note that the produced current—and thus the spin-charge conversion rate—is remanent at gate voltage $V_{\text{gate}} = 0 \text{ V}$, as seen in C and E in Fig. 2a and b. Similar hysteretic behaviours have been obtained on several thinned-down samples but not on a 500- μm -thick STO substrate, which indicates the existence of a critical electric field for the hysteresis to appear. The non-volatile control of the spin-charge conversion is further evidenced by Fig. 2c, which displays the normalized charge current produced at 0 V after the application of 500-ms pulses of $\pm 200 \text{ V}$ gate voltage. Figure 2d shows the temperature dependence of the difference ΔI_C in the produced current obtained at remanence after applying pulses of $+200 \text{ V}$ and -200 V at 7 K . ΔI_C is large below 30 K and vanishes above $45\text{--}50 \text{ K}$, suggesting a transition of STO into the paraelectric phase⁷⁻⁹. Extended Data Figs. 1 and 4 show that the effect is reproducible and stable in time for at least several hours.

We have also performed electric polarization measurements on a Al(1.8 nm)/STO 2DEG sample with a STO thickness of $200 \pm 20 \mu\text{m}$. As visible in Fig. 3a, the application of an electric field up to 2.5 kV cm^{-1} (green curve) yields a linear dependence of the polarization with E , as expected for a dielectric. However, when the voltage exceeds about 7 kV cm^{-1} , a hysteresis develops, associated with switching current peaks in the I versus E data (Fig. 3a inset). The saturation polarization is about $4 \mu\text{C cm}^{-2}$, in agreement with earlier reports⁷. Upon increasing the temperature (Fig. 3c), the loop progressively closes, indicating a Curie temperature close to 50 K (Fig. 3b). This almost coincides with the temperature at which the remanent spin-charge conversion effect vanishes (Fig. 2d), strongly suggesting that the switchable polarization is at the origin of the hysteretic inverse Edelstein effect. At low temperature, reducing E to below the critical value still yields

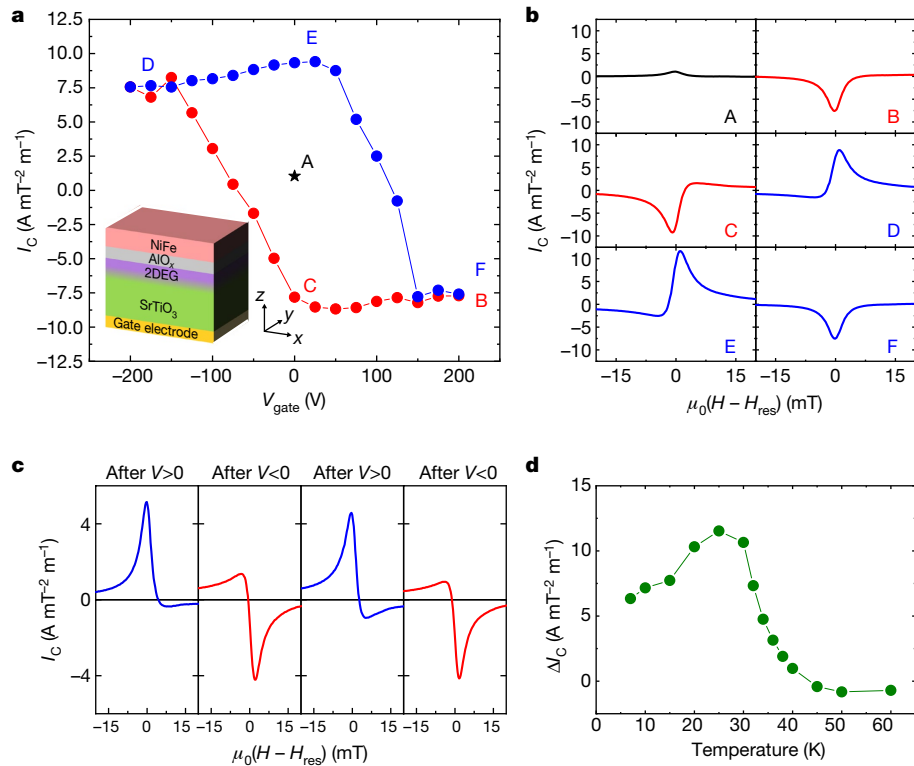


Fig. 2 | Electric-field-controlled spin-charge conversion with electrical remanence. **a**, Gate-voltage dependence of the normalized current produced by the inverse Edelstein effect. A–F, I_C conditions examined in **b**. Inset, a sketch of the heterostructure. **b**, Magnetic-field dependence of the normalized current produced in spin-pumping experiments, for different voltage values

(see **a**). **c**, Normalized charge current produced at electrical remanence after applying positive or negative voltage pulses of ± 200 V. All data have been measured at 7 K. **d**, Temperature dependence of the difference between the remanent normalized currents after the application of a large positive or negative voltage.

hysteretic polarization loops, albeit with a lower remanent polarization (Fig. 4a).

One of the hallmark features of STO 2DEGs is the strong gate-voltage dependence of the sheet resistance²⁹ R_s . In thick STO samples the

gate dependence of R_s is usually non-hysteretic³⁰, in line with the paraelectric nature of STO at low electric fields. Here, as seen in Fig. 4b, R_s varies as the carrier density varies, but this dependence also exhibits a clear hysteresis, allowing the non-volatile electric control

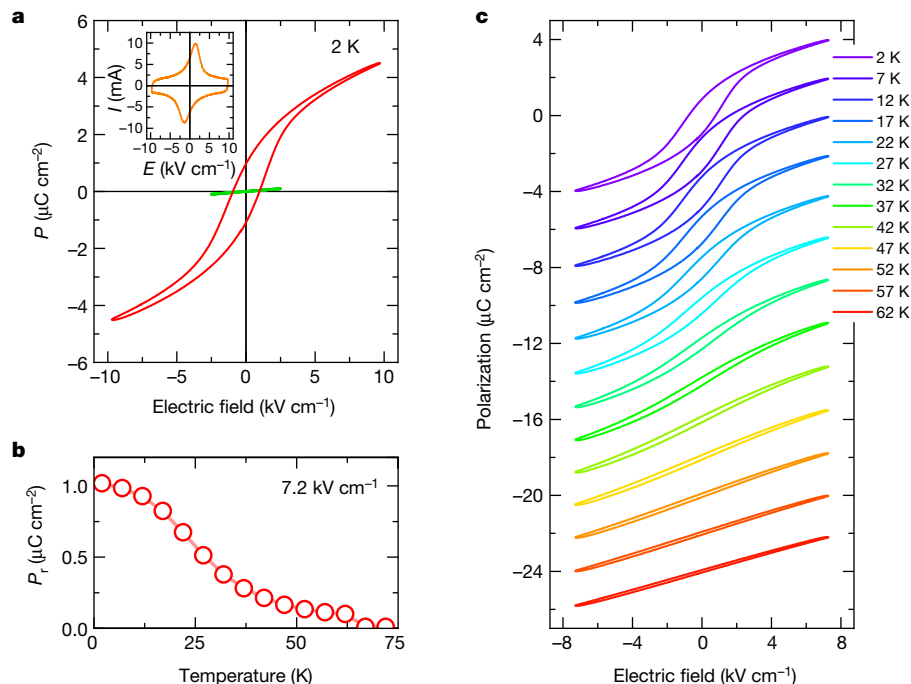


Fig. 3 | Electric polarization measurements. **a**, Polarization versus electric field curves measured on a Al(1.8 nm)/STO sample. The green curve corresponds to the polarization loop measured with a maximum field of

2.5 kV cm⁻¹. Inset, corresponding current versus electric field curve. **b**, Temperature dependence of the remanent polarization P_r . **c**, Polarization loops at different temperatures.

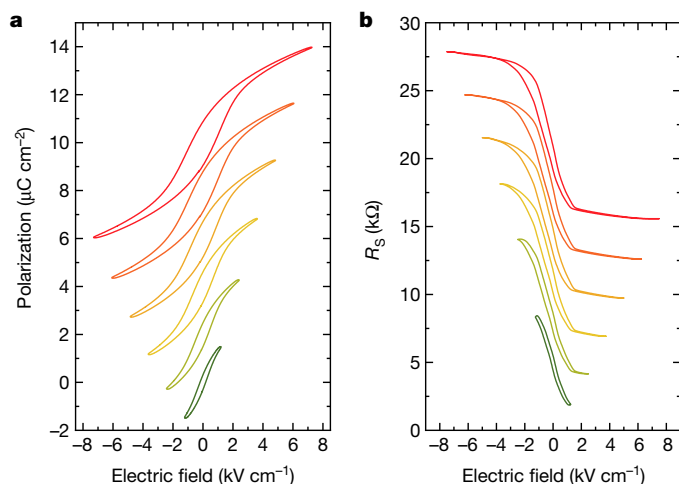


Fig. 4 | Field effect experiments. **a**, Polarization loops at 2 K measured in the field-induced state for different increasing maximum electric fields. The curves are shifted by $2 \mu\text{C cm}^{-2}$ for clarity. **b**, Gate dependence of the 2DEG sheet resistance R_s for different maximum electric fields at 2 K. The curves are shifted by $3 \text{ k}\Omega$ for clarity.

of the 2DEG electronic properties. We note that the hysteresis amplitude increases upon increasing the maximum E , so that the R_s versus E loops mimic the polarization loops of Fig. 4a. Hall measurements made in the two remanent states yield a difference in carrier densities $\Delta n_s = 5.45 \times 10^{12} \text{ cm}^{-2}$, only two times smaller than the theoretical value $\Delta n_s = 2P_r/e = 1.13 \times 10^{13} \text{ cm}^{-2}$ (using the remanent polarization $P_r = 0.9 \mu\text{C cm}^{-2}$); this corresponds to a notable efficiency compared to the literature^{31,32}. Note that we have also performed R_s versus E loops on spin-pumping samples, which possess a NiFe layer, showing that the obtained loops are very similar to the λ_{IEE} versus V_{gate} loops (see Extended Data Fig. 1 and Methods).

Several mechanisms may be invoked to explain our observation of a hysteretic inverse Edelstein effect. One can be related to the description of Fig. 1a, namely, a local inversion of the electric field in the SOC material (here the 2DEG) promoting polarization-direction-dependent Rashba SOC and spin-charge conversion. Additionally, electronic structure effects may be at play, since the multiorbital band structure of STO 2DEGs is known to produce effective Rashba effects with opposite signs, depending on the orbitals involved^{25,26}. Moreover, the presence of a switchable polarization with associated polar displacements of cations and anions should substantially modify the band structure compared to the paraelectric case. This may generate additional (avoided) band crossings, possibly with non-trivial topology²⁵, leading to super-efficient spin-charge conversion.

Our results constitute the basis of a new type of spintronics, in which non-volatility would not originate from ferromagnetism but from ferroelectricity. They could be extended to room temperature by, for example, designing 2DEGs on strained STO thin films³³ or BaTiO₃ (ref. ²⁴). This could open the way to a new class of ultralow-power spin-orbitronic devices (such as memories, spin field-effect transistors, spin Hall transistors or MESO-like logic devices). In the future, demonstration of a non-volatile electric control of the direct Edelstein effect could additionally lead to reconfigurable spin-orbit torque memories and logic gates, be of benefit to the manipulation of skyrmions or domain walls, and allow the development of agile terahertz emitters and spin-wave logic architectures.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at [<https://doi.org/10.1038/s41586-020-2197-9>].

- Jones, N. How to stop data centres from gobbling up the world's electricity. *Nature* **561**, 163–166 (2018).
- Manipatruni, S., Nikonov, D. E. & Young, I. A. Beyond CMOS computing with spin and polarization. *Nat. Phys.* **14**, 338–343 (2018).
- Žutić, I., Fabian, J. & Das Sarma, S. Spintronics: fundamentals and applications. *Rev. Mod. Phys.* **76**, 323–410 (2004).
- Slonczewski, J. C. Current-driven excitation of magnetic multilayers. *J. Magn. Magn. Mater.* **159**, L1–L7 (1996).
- Heron, J. T. et al. Deterministic switching of ferromagnetism at room temperature using an electric field. *Nature* **516**, 370–373 (2014).
- Gränicher, H. Induzierte Ferroelektrizität von SrTiO₃ bei sehr tiefen Temperaturen und über die Kälteerzeugung durch adiabatische Entpolarisierung. *Helv. Phys. Acta* **29**, 210–212 (1956).
- Hemberger, J., Lunkenheimer, P., Viana, R., Böhrer, R. & Loidl, A. Electric-field-dependent dielectric constant and nonlinear susceptibility in SrTiO₃. *Phys. Rev. B* **52**, 13159–13162 (1995).
- Sidoruk, J. et al. Quantitative determination of domain distribution in SrTiO₃ — competing effects of applied electric field and mechanical stress. *J. Phys. Condens. Matter* **22**, 235903 (2010).
- Manaka, H., Nozaki, H. & Miura, Y. Microscopic observation of ferroelectric domains in SrTiO₃ using birefringence imaging techniques under high electric fields. *J. Phys. Soc. Jpn* **86**, 114702 (2017).
- Caviglia, A. D. et al. Tunable Rashba spin-orbit interaction at oxide interfaces. *Phys. Rev. Lett.* **104**, 126803 (2010).
- Ohtomo, A. & Hwang, H. Y. A high-mobility electron gas at the LaAlO₃/SrTiO₃ heterointerface. *Nature* **427**, 423–426 (2004); correction 441, 120 (2006).
- Soumyanarayanan, A., Reyren, N., Fert, A. & Panagopoulos, C. Emergent phenomena induced by spin-orbit coupling at surfaces and interfaces. *Nature* **539**, 509–517 (2016).
- Edelstein, V. M. Spin polarization of conduction electrons induced by electric current in two-dimensional asymmetric electron systems. *Solid State Commun.* **73**, 233–235 (1990).
- Kondou, K. et al. Fermi-level-dependent charge-to-spin current conversion by Dirac surface states of topological insulators. *Nat. Phys.* **12**, 1027–1031 (2016).
- Hoffmann, A. Spin Hall effects in metals. *IEEE Trans. Magn.* **49**, 5172–5193 (2013).
- Bychkov, Y. A. & Rashba, E. I. Properties of a 2D electron gas with lifted spectral degeneracy. *JETP Lett.* **39**, 78–81 (1984).
- Sánchez, J. C. R. et al. Spin-to-charge conversion using Rashba coupling at the interface between non-magnetic materials. *Nat. Commun.* **4**, 2944 (2013).
- Picozzi, S. Ferroelectric Rashba semiconductors as a novel class of multifunctional materials. *Front. Phys.* **2**, <https://doi.org/10.3389/fphy.2014.00010> (2014).
- Rinaldi, C. et al. Ferroelectric control of the spin texture in GeTe. *Nano Lett.* **18**, 2751–2758 (2018).
- Kolobov, A. V. et al. Ferroelectric switching in epitaxial GeTe films. *APL Mater.* **2**, 066101 (2014).
- Rinaldi, C. et al. Evidence for spin to charge conversion in GeTe(111). *APL Mater.* **4**, 032501 (2016).
- Bibes, M., Vila, L., Attané, J.-P., Noël, P. & Vaz, D. C. Dispositif électronique, porte numérique, composant analogique et procédé de génération d'une tension. French patent FR18 74319 (2018).
- Manipatruni, S. et al. Scalable energy-efficient magnetoelectric spin-orbit logic. *Nature* **565**, 35–42 (2019).
- Rödel, T. C. et al. Universal fabrication of 2D electron systems in functional oxides. *Adv. Mater.* **28**, 1976–1980 (2016).
- Vaz, D. C. et al. Mapping spin-charge conversion to the band structure in a topological oxide two-dimensional electron gas. *Nat. Mater.* **18**, 1187–1193 (2019).
- Lesne, E. et al. Highly efficient and tunable spin-to-charge conversion through Rashba coupling at oxide interfaces. *Nat. Mater.* **15**, 1261–1266 (2016).
- Tserkovnyak, Y., Brataas, A. & Bauer, G. E. W. Enhanced Gilbert damping in thin ferromagnetic films. *Phys. Rev. Lett.* **88**, 117601 (2002).
- Noel, P. et al. Highly efficient spin-to-charge current conversion in strained HgTe surface states protected by a HgCdTe layer. *Phys. Rev. Lett.* **120**, 167201 (2018).
- Caviglia, A. D. et al. Electric field control of the LaAlO₃/SrTiO₃ interface ground state. *Nature* **456**, 624–627 (2008).
- Biscaras, J. et al. Limit of the electrostatic doping in two-dimensional electron gases of LaXO₃ (X = Al, Ti)/SrTiO₃. *Sci. Rep.* **4**, 6788 (2015).
- Crassous, A. et al. Nanoscale electrostatic manipulation of magnetic flux quanta in ferroelectric/superconductor BiFeO₃/YBa₂Cu₃O_{7-δ} heterostructures. *Phys. Rev. Lett.* **107**, 247002 (2011).
- Yamada, H. et al. Ferroelectric control of a Mott insulator. *Sci. Rep.* **3**, 2834 (2013).
- Haeni, J. H. et al. Room-temperature ferroelectricity in strained SrTiO₃. *Nature* **430**, 758–761 (2004).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Sample preparation

NiFe and Al films were deposited at room temperature by d.c. magnetron sputtering on TiO₂-terminated (001)-oriented STO substrates (from CrysTec). TiO₂-termination was achieved through a chemical treatment, where the substrate was submerged in a buffered hydrofluoric acid (NH₄F-HF 7:1) for 30 s and annealed under an oxygen-rich environment at 1,000 °C for 3 h. Before deposition, the STO substrates were additionally annealed at 730 °C for 2 h under a partial oxygen pressure of 400 mbar. The deposition of the metallic layers was performed under an Ar partial pressure of 4.5×10^{-4} mbar and a substrate-to-target distance of 7 cm. The samples including NiFe were additionally capped with a 2.5-nm layer of Al, which becomes oxidized when exposed to air. Samples were mechanically polished on diamond pads under deionized water flow.

Spin pumping

The spin-pumping experiments were carried out using a Bruker ESP300E X-band CW spectrometer at 9.68 GHz, with a loop-gap Bruker ER 4118X-MS5 cavity, and using a microwave power of 5 mW or less to remain in the linear regime. The generated d.c. voltage was measured using a Keithley 2182A nanovoltmeter. The gate voltage was applied using a Keithley 2400 sourcemeter. The measured signals were observed to be linear with the r.f. power up to 5 mW.

Calculation of the inverse Edelstein length

The inverse Edelstein length λ_{IEE} is the figure of merit quantifying the efficiency of the spin to charge current conversion. It has the dimension of a length, as the 3D spin current J_S^{3D} (in A m⁻²) is converted into a 2D charge current J_C^{2D} (in A m⁻¹):

$$\lambda_{\text{IEE}} = \frac{J_C^{2D}}{J_S^{3D}} \quad (1)$$

Both J_S^{3D} and J_C^{2D} need to be evaluated to calculate the inverse Edelstein length. Here we use the method already described in previous works (for example, on LAO/STO (ref. ²⁶) or HgTe (ref. ²⁸)).

The produced charge current is simply extracted from the symmetric component of the measured spin signal V_{sym} :

$$J_C^{2D} = \frac{V_{\text{sym}}}{Rw} \quad (2)$$

where R is the resistance of the sample (measured independently), and w is the sample width (400 μm).

Note that here, the produced current J_C^{2D} is used to give the amplitude of the spin signal, as it can be considered as raw data. In order to have values comparable from measurement to measurement, especially with experiments found in the literature, and as the spin signal varies linearly with the square of the excitation field $\mu_0 h_{\text{rf}}$ (h_{rf} is the radio-frequency field), the current production has to be normalized. Thus, the produced current given in the main text is actually $J_C^{2D} / (\mu_0 h_{\text{rf}})^2$, in A mT⁻² m⁻¹. The radiofrequency field for a given measurement is measured using the cavity conversion factor.

The spin current is extracted using the spin-pumping theory first developed by Tserkovnyak, Brataas^{27,34} and co-workers, and then by several other groups^{35,36}. The spin current injected at the ferromagnetic resonance can be obtained by measuring some of the magnetic properties of the ferromagnetic layer, and by calculating the effective spin mixing conductance:

$$G_{\text{eff}}^{\uparrow\downarrow} = \frac{4\pi M_s t_{\text{FM}}}{g\mu_B} (\alpha - \alpha_{\text{ref}}) \quad (3)$$

where μ_B is the Bohr magneton, t_{FM} the thickness of the ferromagnetic material (20 nm here), M_s the saturation magnetization of the Permalloy thin film, g its g -factor, α its Gilbert damping, and α_{ref} the Gilbert damping of a Permalloy thin film without a spin-sink (here Permalloy on native Si). All these values are extracted from independent ferromagnetic resonance (FMR) measurements, using either broadband-FMR or out-of-plane angular dependence measurements.

Then, using the expression of the spin mixing conductance we can obtain the injected spin current:

$$J_S^{3D} = \frac{G_{\text{eff}}^{\uparrow\downarrow} \gamma^2 \hbar \mu_0 h_{\text{rf}}^2}{8\pi\alpha^2} \left[\frac{4\pi M_s \gamma + \sqrt{(4\pi M_s \gamma)^2 + 4\omega^2}}{(4\pi M_s \gamma)^2 + 4\omega^2} \right] \frac{2e}{\hbar} \quad (4)$$

where γ is the gyromagnetic ratio, ω the angular frequency, e the elementary charge and \hbar the reduced Planck constant. The inverse Edelstein length can then be obtained by combining equations (1), (2) and (4).

Gate-voltage dependence reproducibility

We have performed spin-pumping measurements on different samples of NiFe(20 nm)/Al(0.9 nm)//STO at 7 K, see Extended Data Fig. 1. Sample 1 is taken from a first batch, whereas samples 2 and 3 are two different samples from the same second batch. The results shown in the main text have been measured on sample 3. After thin film deposition on STO substrates, the samples were all thinned down to the same thickness (250 ± 20 μm). As can be seen in Extended Data Fig. 1, for these three samples similar gate-voltage dependences have been obtained, with a hysteretic behaviour, a positive or negative remanent spin-signal at $V_{\text{gate}} = 0$ V, and large conversion efficiencies. The obtained inverse Edelstein lengths λ_{IEE} are above 40 nm in all three cases, and up to 60 nm in the case of sample 3. The error bars are mostly due to the uncertainty on the effective spin mixing conductance. The main results presented in the text are thus reproducible, even though the samples have been thinned down using mechanical polishing.

We have also performed several cool-downs on the same sample. After performing a first cool-down and some gate dependence measurements at low temperature, it is possible to recover the initial state by heating up the sample at room temperature. As can be seen in Extended Data Fig. 2 (measured on sample 2), the remanent ferroelectric state is lost after heating, but when going back to 7 K the sample recovers the initial state, with a lower and positive spin signal. This is consistent with our observation of a voltage-induced ferroelectricity at low temperature. After heating at room temperature, an initialization loop [+200 V; -200 V; +200 V] performed at low temperature allows retrieval of the hysteretic behaviour and the remanence of the polarization.

Time stability of the remanent state

In the main text we show that a ± 200 V gate-voltage application at 7 K allows the spin-charge conversion to be controlled in a remanent way. To demonstrate the non-volatility associated with this remanence, we performed spin-pumping measurements hours after applying a gate voltage of either +200 V or -200 V for 500 ms. As seen in Extended Data Fig. 3, the produced normalized current is preserved, remaining unmodified after several hours.

Electric polarization measurements

In these experiments, a triangular waveform was applied at a frequency of 1 kHz across the STO, between the 2DEG and a bottom electrode of Ti/Au, and the current I was measured in real time. Integrating the current over time and normalizing by the sample area yields the polarization.

Magnetotransport

Low-temperature electrical transport measurements were performed on the thinned samples bonded by Al wires in the van der Pauw

configuration using a standard a.c. lock-in technique ($I_{ac} = 200$ nA, $f_{ac} = 77.03$ Hz) in a Quantum Design Dynacool cryostat at a temperature of 2 K and magnetic fields between -9 T and 9 T for the Hall resistance study. Before any back-gate voltage data were recorded, the samples were subjected to a so-called forming step³⁰ at 2 K, where the back-gate voltage were cycled several times (>2) between the gate-voltage extremes of the particular gate-voltage interval to ensure no irreversible changes would occur in the interface system upon application of the back-gate voltage in the actual experiment. Note that this low-temperature forming step was repeated following all occasions the sample was brought above 105 K. Moreover, at each new cool-down, the samples were always cooled with the back-gate electrostatically grounded.

R–V loops measured on NiFe/Al/SrTiO₃ samples

Extended Data Figure 3 shows R – V loops measured on the NiFe/Al/STO sample used for spin pumping. The R – V and I_c – V loops have rather similar shapes, indicating a similar origin for both hysteresis effects. The observed two-probe resistance variation of ~ 0.27 Ω in this 0.4 mm \times 2.4 mm NiFe(20 nm)/AlO_x/STO sample is compatible with the R – V for an AlO_x/STO sample shown in Fig. 4b. The room-temperature sheet resistance of the NiFe(20 nm)/AlO_x/STO sample is roughly that of the NiFe layer, and equal to 9 Ω . In Fig. 4, gating results in a change of the 2DEG sheet resistance from about 1.7 k Ω to 23.5 k Ω . In a simple parallel model of the NiFe(20 nm)/AlO_x/STO sample (in which current flows in parallel in the NiFe and the 2DEG), gating should thus result in a sheet resistance change of

$$\begin{aligned}\Delta R &= \left(\frac{R_S^{2DEG} R_S^{NiFe}}{R_S^{2DEG} + R_S^{NiFe}} \right)_{VG-} - \left(\frac{R_S^{2DEG} R_S^{NiFe}}{R_S^{2DEG} + R_S^{NiFe}} \right)_{VG+} \\ &= \frac{23,500 \times 9}{23,500 + 9} - \frac{1,700 \times 9}{1,700 + 9} \\ &= 0.044 \Omega\end{aligned}$$

corresponding to an expected two-probe resistance change of 0.26 Ω , in excellent agreement with the observed change of 0.27 Ω . $VG-$ and $VG+$ indicate negative and positive gate voltages, respectively.

The shape of the P – V and R – V loops of Fig. 4 is different from that of the I_c – V and R – V data of Extended Data Fig. 3. One reason is that the spin-pumping experiments were performed on a Al/STO sample

covered with a NiFe layer to perform the spin injection, whereas the resistance versus electric field (R – E) and polarization versus electric field (P – E) loops were performed on Al/STO samples without NiFe and thus with a different electrostatic geometry. Additionally, the sample dimensions are also different for the two sets of experiments. In the spin pumping FMR experiments, the STO thickness is 250 μ m, and the lateral size is 0.4 mm \times 2.4 mm. For the R – E and P – E loop experiments, the STO thickness is 200 μ m and the lateral size is 5 mm \times 5 mm. Finally, the SP-FMR samples are cut from plain samples, which could induce defects modifying the coercivity. We believe the observed discrepancy between loops to arise primarily from these above-mentioned differences.

Data availability

The data that support the findings of this study are available from the corresponding authors upon reasonable request.

34. Brataas, A., Tserkovnyak, Y., Bauer, G. E. W. & Halperin, B. I. Spin battery operated by ferromagnetic resonance. *Phys. Rev. B* **66**, 060404 (2002).
35. Costache, M. V., Sladkov, M., Watts, S. M., van der Wal, C. H. & van Wees, B. J. Electrical detection of spin pumping due to the precessing magnetization of a single ferromagnet. *Phys. Rev. Lett.* **97**, 216603 (2006).
36. Ando, K. et al. Inverse spin-Hall effect induced by spin pumping in metallic system. *J. Appl. Phys.* **109**, 103913 (2011).

Acknowledgements The authors thank M. Cazayous, B. Dkhil, M. Maglione, S. Gambarelli and V. Maurel for useful discussions, as well as C. Carrétéro, E. Jacquet and Y. Gourdel for technical help. This work received support from the ERC Consolidator grant number 615759 “MINT”, the ERC Advanced grant number 833973 “FRESCO”, the QUANTERA project “QUANTOX”, the French Research Agency (ANR) as part of the “Investissement d’Avenir” programme (LABEX NanoSaclay, ref. ANR-10-LABX-0035) through project “AXION” and the Laboratoire d’Excellence LANEF (ANR-10-LABX-51-01) and ANR project OISO (ANR-17-CE24-0026-03). F.T. acknowledges support by research grant VKR023371 (SPINOX) from VILLUM FONDEN.

Author contributions J.-P.A., P.N., L.V. and M.B. designed the experiment. J.-P.A., L.V. and M.B. supervised the study. D.C.V., L.M.V.A. and J.B. prepared the samples. P.N. performed the spin-charge conversion experiments with J.-P.A. and L.V. J.B., S.F. and M.B. performed the polarization measurements with the help of V.G. and F.T. F.T. and J.B. performed the transport experiments and analysed them with M.B. and A.B. M.B. and J.-P.A. wrote the paper with inputs from all authors.

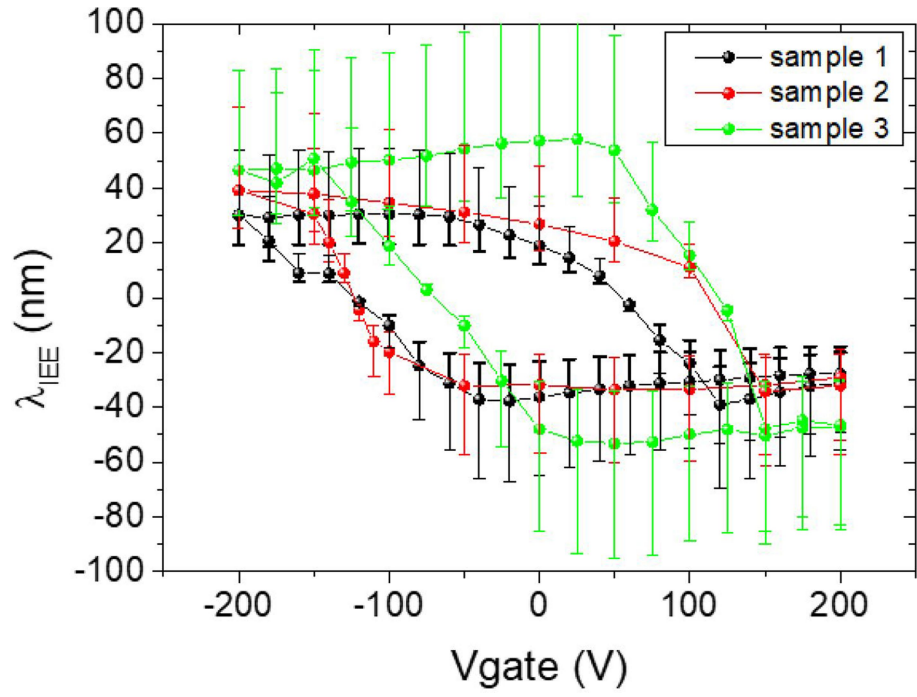
Competing interests The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.B. or J.-P.A.

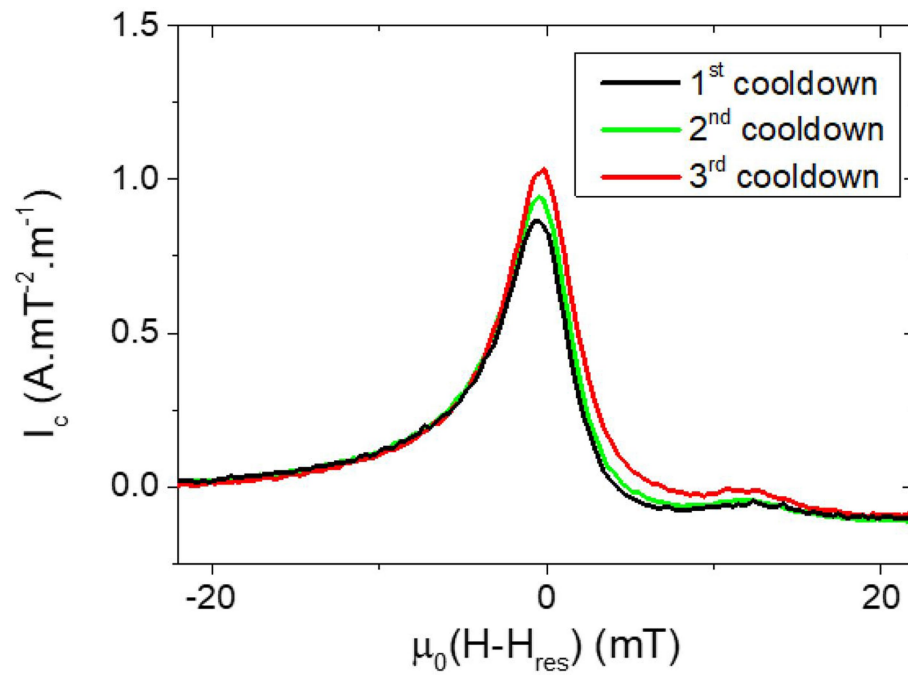
Peer review information Nature thanks Dmitri E. Nikonov, Sashi Satpathy and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

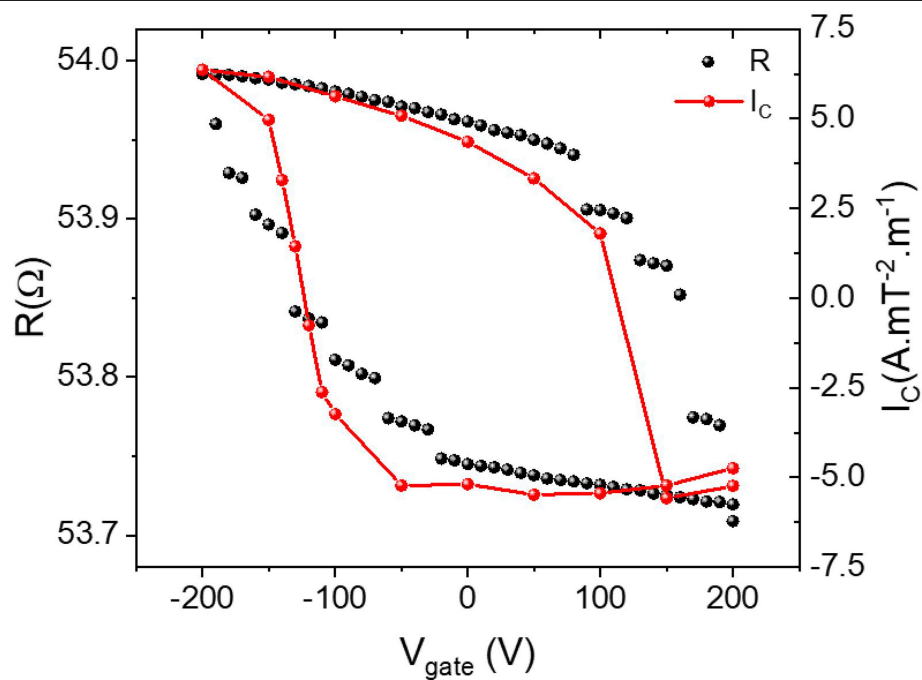


Extended Data Fig. 1 | Gate-voltage dependence of the inverse Edelstein length in three different samples of NiFe(20 nm)/Al(0.9 nm)//STO. The error bars are due to the small extra damping measured in this system. The estimated

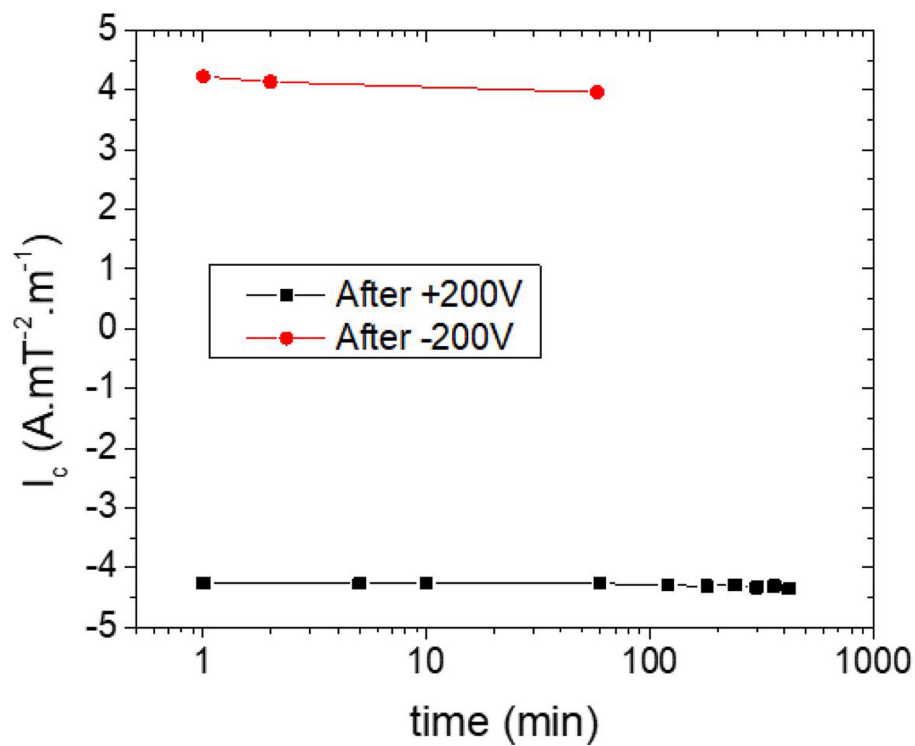
effective spin mixing conductance $G_{\text{eff}}^{\uparrow\downarrow}$ is ranging from 1.2 nm^{-2} to 3.2 nm^{-2} with a mean value of 2.2 nm^{-2} , leading to an injected spin current J_S^{3D} ranging from 100 to $240 \text{ MA m}^{-2} \text{ mT}^{-2}$, with a mean value of $160 \text{ MA m}^{-2} \text{ mT}^{-2}$.



Extended Data Fig. 2 | Spin-pumping signals obtained at 7 K on sample 2, for three different cool-downs from room temperature. After each cool-down, the signal was measured before any gate-voltage application.



Extended Data Fig. 3 | Spin-pumping and resistance loops of a NiFe/Al/STO sample. Black data points, two-probe resistance R of a NiFe/Al/STO sample, measured in the spin-pumping setup as a function of the back-gate voltage. Red data points, normalized charge current production (I_c) measured by spin pumping.



Extended Data Fig. 4 | Dependence of the produced current on the time spent after application of a positive or negative gate voltage. Black squares, +200 V; red circles, -200 V. The measurements were performed at 7 K on sample 1.

Ionic solids from common colloids

<https://doi.org/10.1038/s41586-020-2205-0>

Received: 14 August 2019

Accepted: 26 February 2020

Published online: 22 April 2020

 Check for updates
Theodore Hueckel¹, Glen M. Hocky¹, Jeremie Palacci² & Stefano Sacanna^{1✉}

From rock salt to nanoparticle superlattices, complex structure can emerge from simple building blocks that attract each other through Coulombic forces^{1–4}. On the micrometre scale, however, colloids in water defy the intuitively simple idea of forming crystals from oppositely charged partners, instead forming non-equilibrium structures such as clusters and gels^{5–7}. Although various systems have been engineered to grow binary crystals^{8–11}, native surface charge in aqueous conditions has not been used to assemble crystalline materials. Here we form ionic colloidal crystals in water through an approach that we refer to as polymer-attenuated Coulombic self-assembly. The key to crystallization is the use of a neutral polymer to keep particles separated by well defined distances, allowing us to tune the attractive overlap of electrical double layers, directing particles to disperse, crystallize or become permanently fixed on demand. The nucleation and growth of macroscopic single crystals is demonstrated by using the Debye screening length to fine-tune assembly. Using a variety of colloidal particles and commercial polymers, ionic colloidal crystals isostructural to caesium chloride, sodium chloride, aluminium diboride and K₄C₆₀ are selected according to particle size ratios. Once fixed by simply diluting out solution salts, crystals are pulled out of the water for further manipulation, demonstrating an accurate translation from solution-phase assembly to dried solid structures. In contrast to other assembly approaches, in which particles must be carefully engineered to encode binding information^{12–18}, polymer-attenuated Coulombic self-assembly enables conventional colloids to be used as model colloidal ions, primed for crystallization.

To assemble ionic colloidal crystals in water, we employ a twist to Derjaguin–Landau–Verwey–Overbeek (DLVO) theory¹⁹ that balances the electrostatic attractive force between oppositely charged particles with steric repulsion from well-defined polymer brushes between them. Attraction is provided by overlapping electrical double layers, that is, clouds of oppositely charged ions surrounding charged particles. The thickness of the double layer is characterized by the Debye screening length λ_D , which sets the range of attraction. The polymer brush serves as a particle spacer and its purpose is twofold: first to prevent particles from entering the van der Waals region, and second to set the overlap of the electrical double layers. Particle separation regulates the amplitude of the electrostatic attraction, effectively forming an ionic bond whose strength can be tuned through λ_D . This idea is illustrated in Fig. 1a, where we consider the contribution of the electrostatic potential V_E (dotted) between overlapping oppositely charged double layers and the repulsive potential V_p (dashed) between polymer brushes. As we typically consider particle spacers of thickness larger than 6 nm, particles do not come in close contact and we can neglect van der Waals interactions. V_E is obtained as for DLVO theory, but considering particles of opposite surface potentials ψ_+ and ψ_- . This gives: $\frac{V_E}{k_B T} = 2\pi\epsilon r\psi_+\psi_-\exp(-h/\lambda_D)$, where r is the radius, h is the surface-to-surface particle separation, k_B is the Boltzmann constant, T is temperature and ϵ is the solvent permittivity¹⁹. The repulsion between the polymer brushes, V_p , is given by the Alexander-de Gennes polymer brush model (Methods), built on scaling arguments and predicting

forces similar to more advanced theoretical treatments²⁰. Superposition of the two yields the pair potential V_{EP} between oppositely charged particles (solid lines), which exhibit a local minimum conveniently tuned by varying λ_D relative to the thickness of the polymer spacer L . The depth of the minima corresponds to the bond energy E_b between oppositely charged particles. For $\lambda_D \ll L$, the polymer brush prevents the double layers from overlapping, the electrostatic interaction vanishes and the particles are sterically stabilized. For $\lambda_D \gg L$, the presence of the polymer brush is negligible, the double layers fully overlap, and the particles aggregate. For λ_D comparable to L , repulsion from the brush and electrostatic attraction nicely balance to establish an ionic bond E_b of a few $k_B T$. Here the particle spacers maintain separation between oppositely charged particles while attracting one another electrostatically, providing cohesion with reconfigurability—an indispensable condition for dependable assembly. Qualitatively, the equilibrium distances prove to be close to the contact point between the two brushes, $h \approx 2L$, as for a nearly impenetrable wall, which provides an intuitive handle to both rationalize the interaction and streamline the experimentation.

Following this simple principle, the assembly of an ionic colloidal crystal requires only three ingredients: (1) oppositely charged particles, (2) a non-ionic surfactant that forms a uniform brush onto the surface of the particles and (3) salt to set the range λ_D of the electrostatic interactions. In a prototypical self-assembly experiment, a block copolymer (Pluronic F108) adsorbs onto a set of oppositely charged polystyrene (PS) microspheres ($r \approx 500$ nm), forming a brush.

¹Department of Chemistry, New York University, New York, NY, USA. ²Department of Physics, University of California San Diego, La Jolla, CA, USA. ✉e-mail: s.sacanna@nyu.edu

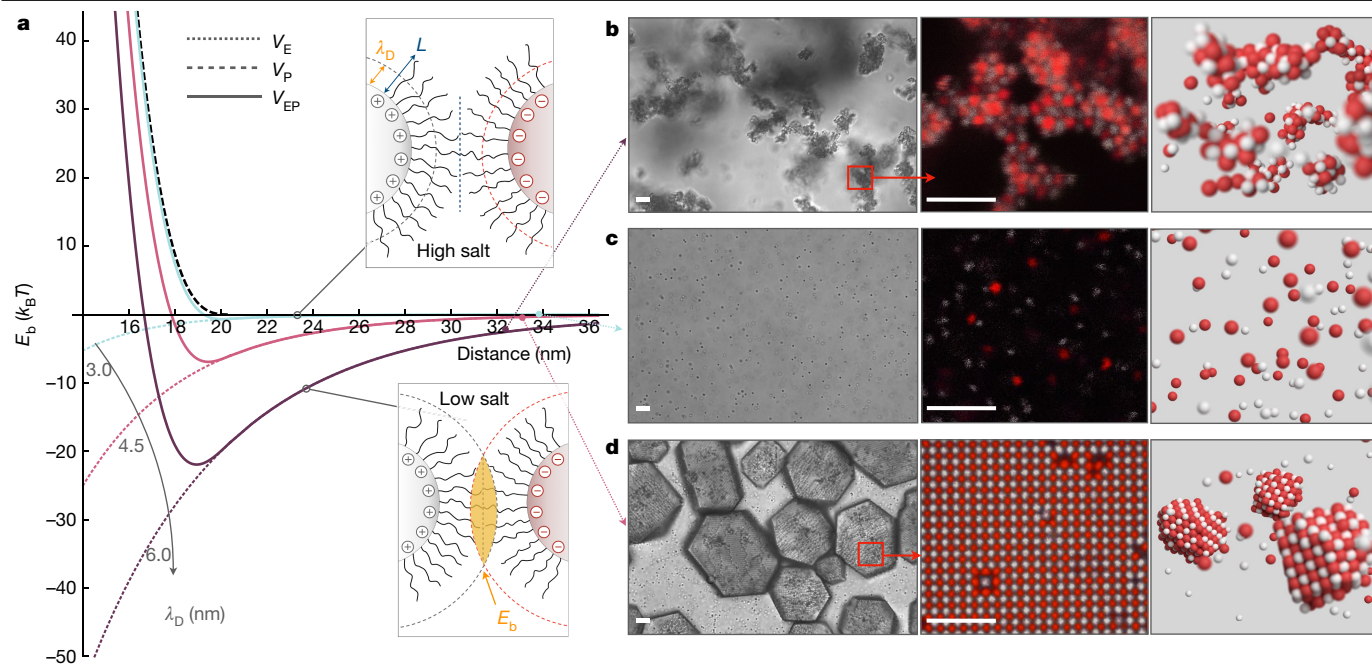


Fig. 1 | Polymer-attenuated Coulombic self-assembly. **a**, Oppositely charged particles separated by a polymer spacer form a tunable ionic bond. Their pair potential (V_{EP}) comprises polymer repulsion (V_P) and Coulombic attraction (V_E). Plots are calculated for a constant polymer brush length $L = 10$ nm and increasing Debye screening lengths λ_D of 3.0, 4.5 and 6.0 nm. For each

value of λ_D , the corresponding V_{EP} minima represent the ionic bond strength E_b . **b–d**, From left to right: bright field microscopy, confocal microscopy and computer simulations of oppositely charged F108-grafted PS spheres at $\lambda_D = 6.0$ (b), 3.0 (c) and 4.5 (d). Scale bars, 4 μm .

The colloids are then equilibrated in sodium chloride (NaCl) solutions and simply mixed together. After mixing, the fate of the binary suspension is set by the salt concentration in the system. We observe three distinct assembly behaviours that are consistent with the relative bond energies calculated using our model. At low salt concentrations (≤ 4.3 mM), E_b is much larger than the thermal energy $k_B T$, causing the particles to bind irreversibly. What follows is a catastrophic flocculation that yields the macroscopic heteroaggregates shown in Fig. 1b. Increasing the salt concentration (≥ 4.9 mM) reduces E_b , leading to a less intuitive scenario whereby oppositely charged particles coexist in a stable suspension (Fig. 1c). In between these two regimes, we find

a narrow window of salinities in which oppositely charged particles behave as model ions, self-assembling into perfectly ordered ionic solids (Fig. 1d). The emergence of vibrant structural colours within a sample (Extended Data Fig. 1, Supplementary Video 1) easily reveals this ‘Goldilocks zone’, which—assuming a brush length $L = 10$ nm for Pluronic F108—is characterized by bond energies of about $8k_B T$. When left undisturbed, the oppositely charged particles condense within hours from mixing, forming millimetre-sized crystalline solids in a matter of days. Particles with stronger attraction initially produce gels that can anneal over time into crystalline phases (Extended Data Fig. 2). Although the resultant crystal domains are smaller than the ideal case,

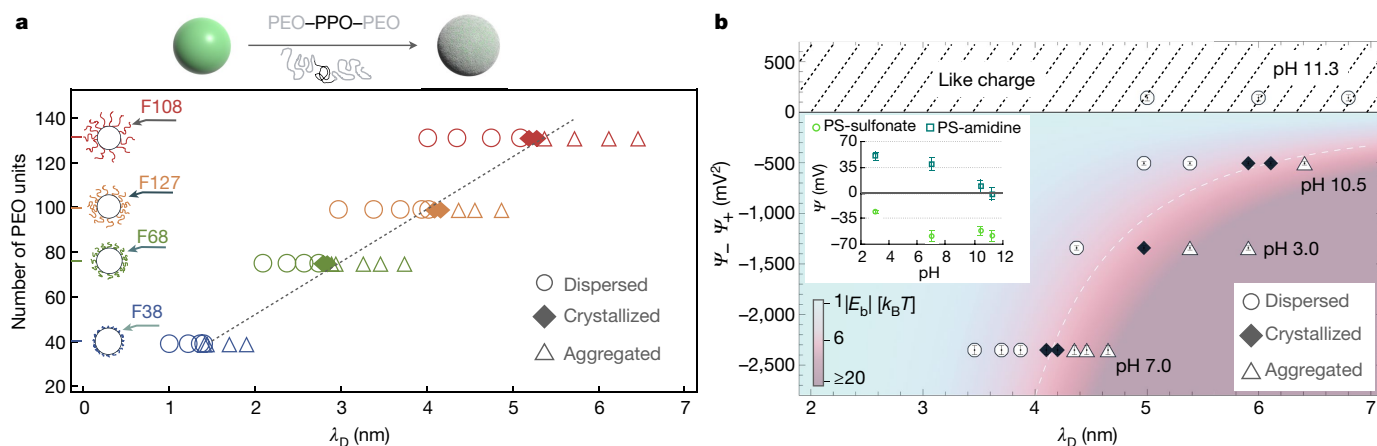


Fig. 2 | Tunable crystallization conditions. **a**, Experimental phase diagram showing the assembly behaviour of oppositely charged PS as a function of λ_D for different brush lengths. Polymer brushes are installed by grafting or physisorption of PEO–PPO–PEO triblock copolymers. The polymer architecture comprises an anchoring block (PPO) and a variable spacer block (PEO). Error bars (± 1 s.d.) of the data points are smaller than the marker size. The dotted line is a linear fit between crystalline points. **b**, Assembly behaviour

of oppositely charged F108–PS colloids as a function of λ_D for different pH values. The colour gradient corresponds to the ionic bond energy E_b calculated assuming a brush length $L = 10$ nm. The pH affects the surface potentials (ψ_+ and ψ_- , inset) of the particles, which in turn set the amplitude of the electrostatic attraction. Error bars are ± 1 s.d. Crystallization occurs along a constant energy line (dotted) across a broad range of pH values. When the opposite charge is lost (here at pH 11.3) the suspension remains dispersed.

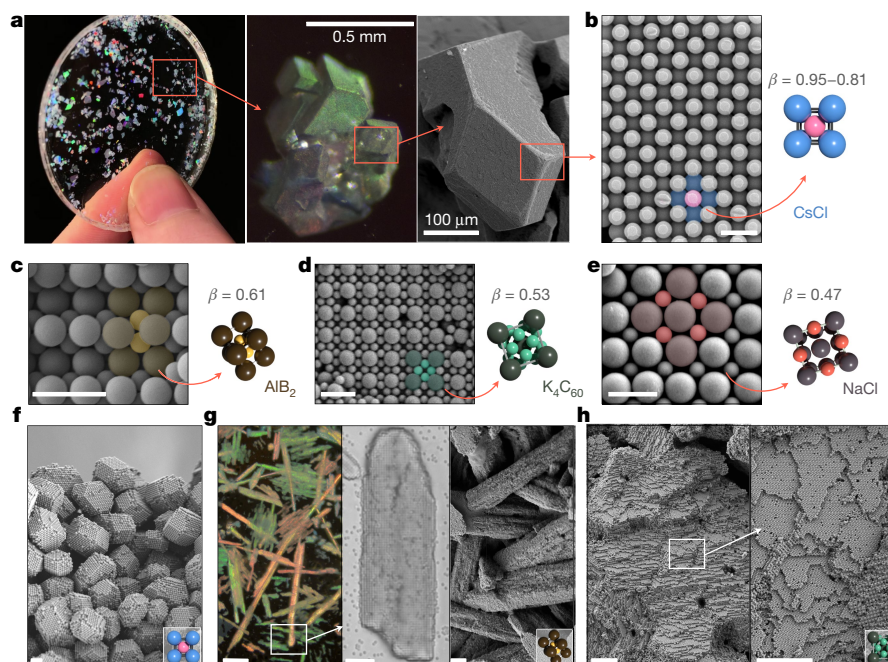


Fig. 3 | Fixed crystals. **a**, Fixed ionic colloidal crystals can be safely dried and manipulated. Left, a collection of millimetre-sized iridescent monocrystalline solids embedded in a matching refractive index epoxy resin (Supplementary Video 4). Middle and right, optical (middle) and SEM (right) micrographs showing the characteristic rhombic dodecahedral habit of a fixed CsCl crystal. **b–e**, Electron microscopy analysis reveals the exact structure of a series of ionic crystals formed at different β values, including CsCl (**b**), AlB_2 (**c**), K_4C_{60} (**d**)

and NaCl (**e**). Scale bars, 1 μm . **f**, SEM image showing rhombic dodecahedral CsCl crystallites. Scale bar, 3 μm . **g**, Left, optical micrograph of needle-like AlB_2 crystals imaged through crossed polarizers. Scale bar, 100 μm . Middle, bright-field micrograph showing a growing AlB_2 needle. Scale bar, 5 μm . Right, SEM image of fixed AlB_2 needles. Scale bar, 20 μm . **h**, SEM images of an exfoliated K_4C_{60} bulk crystal. Scale bars, 10 μm .

this demonstrates that even when particles seem to aggregate chaotically, they remain sterically stabilized and continue to reconfigure, albeit slowly. In addition to direct visual clues, the nucleation and growth of ionic solids can be followed in much greater detail by in situ confocal microscopy, which allows differentiation between positive and negative species using two different fluorescent dyes (Fig. 1b–d, centre). Molecular dynamics simulations (Methods) further support our simple interaction model by revealing assembly behaviour that closely matches the experimental observations (Fig. 1b–d, right, Supplementary Video 2).

We perform crystallization experiments varying every ingredient in the system, including type of particle spacer, solvent conditions and building blocks. First, we explore the polymer-attenuated Coulombic self-assembly (PACS) energy landscape by systematically changing the length of the spacers. We select various polymer brush thicknesses through a series of Pluronic surfactants—namely F108, F127, F68 and F38, listed in decreasing length L . For each spacer, we rationally select the screening length to reach E_b of approximately $8k_bT$ based on calculations from our model, and then fine-tune the experimental conditions until crystals nucleate. As expected, particles with shorter spacers require higher salt concentrations (that is, smaller values of λ_D) to reproduce the same self-assembly behaviour observed for longer spacers. We found a direct correlation between the spacer length and the values of λ_D that cause crystallization, such that a constant ratio between the two produces equivalent interaction strengths (Fig. 2a). This confirms the simple argument that well-defined separation between oppositely charged particles allows for tunable electrostatic assembly. The linear relationship is clear across the three longest polymers, whereas the shortest polymer F38 fails to produce crystals, which we attribute to displacement flocculation because of its weak anchoring. We further test the robustness of PACS by assembling crystals under different pH conditions, as this strongly affects the native charge of the colloids.

Figure 2b shows that as we move away from neutral pH, one particle type rapidly loses charge—either the positive in basic conditions or the negative in acidic conditions. As the charge vanishes, the amplitude of the electrostatic attraction decreases, causing the crystals to disassemble. Increasing values of λ_D can compensate for the electrostatic decay, thus resulting in suspensions that crystallize at lower salt concentrations. By applying this criterion, crystals assemble over a broad range of pH conditions, with limits that are set only by the isoelectric points of the particles, beyond which the opposite charge is lost and electrostatic attraction vanishes. Finally, we mix and match particles of different compositions to demonstrate that PACS is not limited to PS microspheres, but applies to virtually any water-based colloidal system. Silica, 3-(trimethoxysilyl) propyl methacrylate (TPM) and PS particles with sizes ranging from 200 nm to 2 μm all successfully crystallize. In particular, the versatility of the method is best illustrated by the formation of hybrid crystals that incorporate both solid and liquid components (Supplementary Video 3).

The carefully balanced interactions that are required for colloidal self-assembly are often incompatible with any chemical or mechanical perturbation, requiring specific fixing mechanisms, ultimately limiting yield and scope^{21,22}. By contrast, PACS interactions establish interparticle bonds with a convenient and general self-locking mechanism that bypasses this issue entirely. This distinctive feature is easily understood by considering the effect of removing salt on the pair potential of the particles. Because of the simple relationship between λ_D and bond strength, crystals can be first assembled in salty water and then allowed to gradually harden by dilution or dialysis. In particular, we find that under deionized conditions, crystals become fixed solids, at which point they can be handled in solution and dried while retaining crystalline order. This process is irreversible, and crystals do not disassemble if salt is added back to the suspension, indicating permanent binding due to van der Waals interactions. Through this self-locking

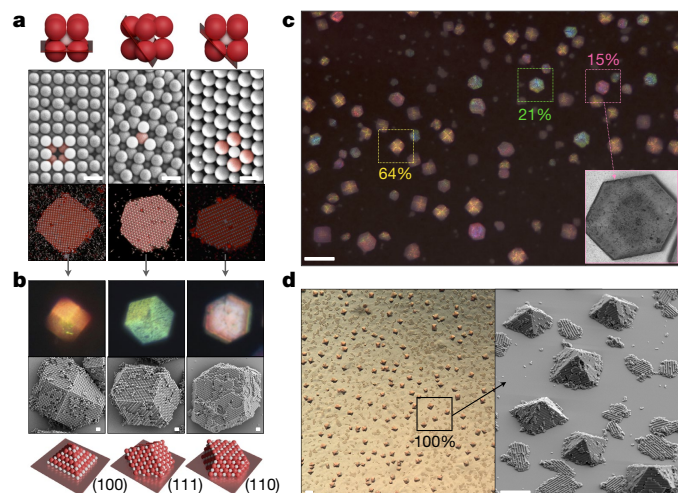


Fig. 4 | Heterogeneous nucleation of crystals. **a**, CsCl crystals growing on a negatively charged glass substrate. From left to right: (100), (111) and (110) planes are imaged by SEM (top) and confocal microscopy (bottom). False-coloured white and red particles highlight positive and negative species, respectively. Scale bars, 1 μm . On top of the SEM images, the schematics illustrate each plane relative to the unit cell of the crystal. **b**, Each crystal plane templates the growth of a macroscopic crystal with a characteristic shape. Crystals are captured by optical microscopy (top) and SEM (bottom). Scale bars, 1 μm . Renderings show the three-dimensional arrangement of the colloidal ions. **c**, Optical microscopy image showing crystals growing on a negatively charged glass substrate and the relative abundance of each crystal shape. Negative substrates favour the nucleation of crystal planes with a high planar density of positive particles. Scale bar, 40 μm . The inset shows the regular shape of a single crystal imaged by bright-field microscopy. **d**, An increased particle–substrate bond energy leads to the nucleation of only (100) planes. Left, bright-field micrograph of monodisperse pyramid-like crystals. Right, an SEM image of the fixed crystals, fully revealing their three-dimensional microstructure. Scale bars, 20 μm .

mechanism, we are able to observe that macroscopic colloidal ionic solids grow as single crystals and develop characteristic habits that resemble those of their atomic counterparts (Fig. 3, Extended Data Fig. 3). Dried products can be transferred to new media, such as matching refractive index liquids or flexible epoxy resins (Fig. 3a, Supplementary Video 4).

The structure of an atomic crystal varies according to the size of the constituent ions. Similarly, the structure of our ionic colloidal crystals is determined by the size ratio (β) of the building blocks. For β between 0.95 and 0.81 colloidal crystals isostructural to caesium chloride (CsCl) nucleate, which rapidly develop with a rhombic dodecahedral habit (Fig. 3b, f), composed of faces from the (110) plane. At lower β values, the next structure observed occurs at $\beta = 0.61$, forming the aluminium diboride (AlB_2) crystal structure, which develops with a characteristic needle-like habit²³ (Fig. 3c, g). At $\beta = 0.53$, an exotic K_4C_{60} phase is observed, which, when fractured, reveals large sheets of its (110) planes (Fig. 3d, h, Extended Data Fig. 4). Finally, $\beta = 0.47$ results in the nucleation of crystals with a familiar NaCl structure (Fig. 3e).

In addition to homogeneous nucleation from bulk suspensions, crystals can readily assemble via heterogeneous nucleation against charged substrates. This second self-assembly route allows us to bias the crystal growth along specific crystallographic directions, effectively shaping the growing solids. This is illustrated in Fig. 4a (see also Extended Data

Fig. 5), in which CsCl planes are selected by a negatively charged glass substrate—namely the (100), (111) and (110) planes—in descending order of planar density of positive particles. Each plane templates the growth of CsCl crystals with a specific orientation, resulting in their speciation into three distinct crystal types with characteristic colours and shapes; yellow squares (100), green hexagons (111) and pink hexagons (110) (Fig. 4b). Specific colouration in the species arises from the uniformly oriented crystals scattering light at the same angle. Counting the species based on colour and shape allows rapid determination of the relative yields, where substantially more (100) planes are nucleated due to their higher density of positive particles, followed logically by the (111) and (110) planes. Substrate influence is more pronounced when the interaction strength is increased, leading to the formation of purely (100) planes. We conclude that differing crystalline pyramids can be selected to self-assemble by surface attraction in experiments and simulations (Fig. 4c, Extended Data Fig. 5).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2205-0>.

- Pauling, L. The principles determining the structure of complex ionic crystals. *J. Am. Chem. Soc.* **51**, 1010–1026 (1929).
- Shevchenko, E. V. et al. Structural diversity in binary nanoparticle superlattices. *Nature* **439**, 55–59 (2006).
- Kalsin, A. M. et al. Electrostatic self-assembly of binary nanoparticle crystals with a diamond-like lattice. *Science* **312**, 420–424 (2006).
- Grzybowski, B. A. et al. Electrostatic self-assembly of macroscopic crystals using contact electrification. *Nat. Mater.* **2**, 241–245 (2003).
- Go, D. et al. Programmable co-assembly of oppositely charged microgels. *Soft Matter* **10**, 8060–8065 (2014).
- Månsson, L. K. et al. Preparation of colloidal molecules with temperature-tunable interactions from oppositely charged microgel spheres. *Soft Matter* **15**, 8512–8524 (2019).
- Mihut, A. M. et al. Assembling oppositely charged lock and key responsive colloids: a mesoscale analog of adaptive chemistry. *Sci. Adv.* **3**, e1700321 (2017).
- Wang, Y. et al. Crystallization of DNA-coated colloids. *Nat. Commun.* **6**, 7253 (2015).
- Leunissen, M. E. et al. Ionic colloidal crystals of oppositely charged particles. *Nature* **437**, 235–240 (2005).
- Bartlett, P. & Campbell, A. I. Three-dimensional binary superlattices of oppositely charged colloids. *Phys. Rev. Lett.* **95**, 128302 (2005).
- Bartlett, P., Ottewill, R. H. & Pusey, P. N. Superlattice formation in binary-mixtures of hard-sphere colloids. *Phys. Rev. Lett.* **68**, 3801–3804 (1992).
- Pusey, P. N. & Vanmegen, W. Phase-behavior of concentrated suspensions of nearly hard colloidal spheres. *Nature* **320**, 340–342 (1986).
- Sacanna, S. et al. Lock and key colloids. *Nature* **464**, 575–578 (2010).
- de Nijs, B. et al. Entropy-driven formation of large icosahedral colloidal clusters by spherical confinement. *Nat. Mater.* **14**, 56–60 (2015).
- Harper, E. S., van Anders, G. & Glotzer, S. C. The entropic bond in colloidal crystals. *Proc. Natl Acad. Sci. USA* **116**, 16703–16710 (2019).
- Nykypanchuk, D. et al. DNA-guided crystallization of colloidal nanoparticles. *Nature* **451**, 549–552 (2008).
- Wang, Y. F. et al. Patchy particle self-assembly via metal coordination. *J. Am. Chem. Soc.* **135**, 14064–14067 (2013).
- Rogers, W. B., Shih, W. M. & Manoharan, V. N. Using DNA to program the self-assembly of colloidal nanoparticles and microparticles. *Nat. Rev. Mater.* **1**, 16008 (2016).
- Hunter, R. J. *Foundations of Colloid Science* 2nd edn (Oxford Univ. Press, 2001).
- Milner, S. T., Witten, T. A. & Cates, M. E. Theory of the grafted polymer brush. *Macromolecules* **21**, 2610–2619 (1988).
- Vutukuri, H. R. et al. Bonding assembled colloids without loss of colloidal stability. *Adv. Mater.* **24**, 412–416 (2012).
- McGinley, J. T. et al. Crystal-templated colloidal clusters exhibit directional DNA interactions. *ACS Nano* **9**, 10817–10825 (2015).
- Seo, S. E. et al. Non-equilibrium anisotropic colloidal single crystal growth with DNA. *Nat. Commun.* **9**, 4558 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

PACS model

We estimate the repulsive potential V_p in Fig. 1a using the Alexander-de Gennes polymer brush model^{24–26}:

$$\frac{V_p}{k_B T} = \frac{16\pi r^2 \sigma^{\frac{3}{2}}}{35} \left[28 \left(\left(\frac{2L}{h} \right)^{\frac{1}{4}} - 1 \right) + \frac{20}{11} \left(1 - \left(\frac{h}{2L} \right)^{\frac{11}{4}} \right) + 12 \left(\frac{h}{2L} - 1 \right) \right], \quad (1)$$

$$0 \leq h \leq 2L$$

where the PS radius r is measured by electron microscopy ($r_{SEM} = 250$ nm), the thickness of the F108 brush L is taken equal to 10 nm (ref. ²⁷) and the polymer surface density σ is estimated as 0.09 poly(ethylene oxide) (PEO) chains per nm² (ref. ²⁸). The attractive electrostatic potential V_E is obtained by approximating the surface potentials of the particles (Ψ_s and Ψ_p) with the measured zeta potentials (+39 mV and –59 mV, respectively).

Colloidal model systems

The principal model systems consist of PS particles synthesized via surfactant-free emulsion polymerization. For example, 600-nm-diameter positive spheres were made from a single-step reaction comprising 500 ml of deionized water, 50 ml of styrene monomer ($\geq 99\%$ from MilliporeSigma) and 0.5 g of the radical initiator 2,2'-azobis(2-methylpropionamide) dihydrochloride (97% from MilliporeSigma). All the ingredients were mixed in a three-neck round-bottom flask, heated to 60 °C and stirred at 330 rpm under nitrogen overnight. After the reaction had concluded, the particles were washed via repeated sedimentation and resuspension cycles. Control over the PS size was achieved by varying the monomer concentration, reaction temperature or reaction time. Negatively charged PS particles were produced in the same fashion, replacing 2,2'-azobis(2-methylpropionamide) dihydrochloride with an equivalent weight amount of potassium persulfate ($\geq 99\%$ from MilliporeSigma). Fluorescent labelling of the PS systems was achieved by a swell–deswell method. In brief, fluorescent dyes dissolved in tetrahydrofuran (THF) were added to particles stabilized with Pluronic F108 (MilliporeSigma) to a final concentration of 30% v/v THF, then diluted by a factor of five before washing the particles via multiple sedimentation and resuspension cycles to set them in pure water. The fluorescent dyes used were rhodamine-labelled aminostyrene and 7-nitrobenzo-2-oxa-1,3-diazole-2-(methylamino)ethanol. Rhodamine-labelled aminostyrene was prepared by adding 48 mg of rhodamine B isothiocyanate (mixed isomers from MilliporeSigma), 12 mg of aminostyrene (97% from MilliporeSigma) and 0.1 ml of tetramethylammonium hydroxide (25 wt% in water from MilliporeSigma) to 200-proof ethanol (10 ml), and stirring the mixture overnight, which was then stored at 5 °C. Monodispersed oil droplets used in the assembly of liquid–solid composite crystals (Supplementary Video 3) were prepared through the condensation of 3-(trimethoxysilyl) propyl methacrylate (TPM, $\geq 98\%$ from MilliporeSigma), by adding 80 μ l of ammonia (28 wt%) to 320 ml of deionized water, followed by the addition of 600 μ l of TPM. This mixture was magnetically stirred for 1 h, which resulted in the nucleation and growth of droplets of approximately 1 μ m in diameter. Fluorescent labelling was achieved with rhodamine B isothiocyanate, coupled to 3-aminopropyl trimethoxysilane for covalent binding to the TPM network. The droplets can be used directly or solidified via a radical polymerization. To use the liquid droplets directly, the emulsion was gently centrifuged, the supernatant removed and the droplets resuspended in deionized water. This wash cycle was repeated three times. To solidify the TPM particles, 20 mg of azobisisobutyronitrile (98% from MilliporeSigma) was added to the suspensions and gently stirred for 5 min; then the suspension was left in a sealed container at 80 °C for 4 h. The particles were then washed in the manner described above.

Self-assembly of ionic colloidal crystals

Positive and negative PS spheres were separately equilibrated in a solution containing 0.03 mM of Pluronic F108 and the desired amount of NaCl, typically 3–5 mM. After 1 h, the suspensions were rapidly mixed together while vortexing and then allowed to crystallize undisturbed in various containers (for example, glass vials, NMR tubes, Eppendorf centrifuge tubes or Petri dishes). Occasionally, samples were sealed in glass capillaries (VitroCom) and monitored over time via optical microscopy. Capillaries were pretreated with a hydrophobizing agent to facilitate the formation of a polymer brush when exposed to Pluronic solutions. The typical hydrophobization protocol consisted of a 20-min exposure to trichloromethyl silane vapour inside of a moisture-free sealed chamber.

Assembly experiments involving particles with different brush lengths were performed using PS with grafted polymers. While not necessary for the successful assembly of crystals, grafting ensures that particles maintain a fully saturated surface regardless of polymer solubility. This allows for a fair comparison between particles that carry different polymer spacers. Pluronic surfactants were permanently grafted to the surface of PS particles via a swell–deswell method²⁹. The specific surfactants, namely, F108, F127 and F68, were added to the PS suspensions at their critical micelle concentration (3.4 mM, 0.8 mM and 20 mM, respectively), then THF was added to the suspension to a final concentration of 50% v/v. The suspensions were left to equilibrate for 1 h, and the THF was diluted to below 5% v/v before washing the particles via centrifugation and resuspension. A polymer solution at the critical micelle concentration was used for the first wash until the THF was reduced to below 1% v/v, then particles were finally set in pure water after three more wash cycles. Pluronic F38 was not grafted to the particles because it failed to produce crystals even when present in high concentrations (~3 mM). We believe that F38 is too small to serve as an effective spacer for our PS model system, so it was not further investigated. The assembly of polymer-grafted PS followed a similar protocol as for the assembly of PS with physisorbed polymers. Typically, two PS suspensions (for example, 3 wt% 550-nm-diameter negative PS and 1.6 wt% 450-nm-diameter positive PS) with the same grafted polymer were equilibrated for 30 min in an excess of that polymer at a concentration of 80 μ M and at a variety of salt concentrations based on the desired λ_D . After equilibration, the oppositely charged suspensions were rapidly mixed and immediately sealed in hydrophobized capillaries using hot wax. The capillaries were left undisturbed for 12 h, at which time the assembly behaviour was observed via optical microscopy.

Our assembly strategy was successful with different types of polymers as long as they provided adequate spacing. The triblock copolymers described above have an A–B–A architecture, PEO–PPO–PEO, in which the central PPO block serves as the polymer anchor. We assembled ionic colloidal crystals, however, also using the polymer Brij s100, which has an A–B structure comprising a 100-unit PEO tail and a stearyl head group. Notably, crystallization occurs at an equivalent salt concentration to Pluronic F127, which has equivalent length PEO chains, suggesting Brij forms a polymer shell of similar thickness.

Fixing crystals

Bulk samples have their supernatant slowly diluted with deionized water, avoiding large-scale flows and shear that would affect the sedimented crystal. The supernatant can be carefully exchanged, removing all the salt, at which point crystals are robust enough to be resuspended and dried. Crystals inside of sealed capillaries are fixed by submerging the capillary in deionized water, and breaking the capillary's ends, taking care to not disturb the product as much as possible. The sample is allowed to equilibrate overnight, then removed to dry slowly for another day. The capillary can then be opened by scoring the sides with a glass cutter and cleaving off the top, exposing the fixed crystalline product. Epoxy resin (Norland 73) can be poured directly onto fixed

dried crystals. Bubbles can form as the resin permeates through the crystal, so the infusion was allowed to proceed for typically 15 min to allow air bubbles to evacuate. The resin was then polymerized through 5 min of ultraviolet light exposure.

Imaging and characterization

Optical images and videos were acquired using a Leica DMI3000 inverted microscope equipped with differential interference contrast optics and a high-resolution camera (Hamamatsu ORCA Flash4.0 sCMOS). Assembled crystals were typically imaged through crossed polarizers using a Nikon D5300 camera optically coupled to the microscope. Fixed crystals were imaged by electron microscopy using a MERLIN (Carl Zeiss) field emission SEM. Fluorescent images were taken using a Leica SP8 confocal microscope. Zeta potentials were measured using a Malvern Zetasizer Nano ZS.

Computer simulations

Simulations were performed using HOOMD-blue v2.5 (compiled in single precision)³⁰, using a single graphics processing unit. The pair interaction between two particles of types N (negative) and P (positive) with radius r_N and r_P was defined by adding V_E to the potentials given by equation (1), using HOOMD-blue's tabulated potential option (with 1,000 interpolation points between the close touching distance, $r_N + r_P$, and $r_N + r_P + 20\lambda_D$). For the steric repulsion term, we use a brush length $L = 10$ nm, and a brush density $\sigma = 0.09$ nm⁻². For the electrostatics, we used surface potentials $\Psi_- = -50$ mV and $\Psi_+ = +50$ mV, a dielectric constant of 80 and a mixing rule for the pre-factor $r = 2/(1/r_N + 1/r_P)$. Unless otherwise stated, simulations were initiated as having 8,000 spherical particles on a simple cubic lattice such that the packing fraction is ϕ , with random assignments of particle type. Values of ϕ of either 0.001 or 0.003 tended to give a good amount of nucleation while still leaving enough free particles to assemble some crystals. Heterogeneous nucleation on a charged surface was modelled by adding an attractive Lennard-Jones wall for N (P) particles in the centre of the simulation box pointing up and down, with Lennard-Jones parameters $\sigma = 2r_{N(P)}$, and a tunable ϵ . For Extended Data Fig. 2e, we also added a repulsive potential to the other particle of type P (N) with the same parameters

but where the potential was shifted up and cut off at its minimum distance $r = 2^{1/6}\sigma_{P(N)}$.

Data availability

The data that support the findings of this study are available from the corresponding author on request.

24. Alexander, S. Polymer adsorption on small spheres—scaling approach. *J. Phys. (Paris)* **38**, 977–981 (1977).
25. Gennes, P. G. D. Stabilité de films polymère/solvant. *C. R. Acad. Sci. II* **300**, 839–843 (1985).
26. Kleshchanok, D., Tuinier, R. & Lang, P. R. Direct measurements of polymer-induced forces. *J. Phys. Condens. Matter* **20**, 073101 (2008).
27. Barnes, T. J. & Prestidge, C. A. PEO–PPO–PEO block copolymers at the emulsion droplet–water interface. *Langmuir* **16**, 4116–4121 (2000).
28. Stenkamp, V. S. & Berg, J. C. The role of long tails in steric stabilization and hydrodynamic layer thickness. *Langmuir* **13**, 3827–3832 (1997).
29. Kim, A. J., Manoharan, V. N. & Crocker, J. C. Swelling-based method for preparing stable, functionalized polymer colloids. *J. Am. Chem. Soc.* **127**, 1592–1593 (2005).
30. Glaser, J. et al. Strong scaling of general-purpose molecular dynamics simulations on GPUs. *Comput. Phys. Commun.* **192**, 97–107 (2015).

Acknowledgements This work was supported primarily by the NSF CAREER award DMR-1653465. We are grateful for shared facilities provided through the Materials Research Science and Engineering Center (MRSEC) and MRI programmes of the National Science Foundation under award numbers DMR-1420073 and DMR-0923251. Computational resources were provided by New York University High Performance Computing. J.P. thanks the Sloan Foundation for support through grant FG-2017-9392 and the National Science Foundation under grant number DMR-1554724. We thank H. Kanwal for assistance in self-assembly experiments and H. Chun for comments on the manuscript and discussions.

Author contributions S.S. led the research. T.H. and S.S. conceived the PACS idea. T.H. synthesized the colloidal systems, and designed and performed the crystallization experiments. G.M.H. designed, performed and analysed the simulations, with input from T.H. and S.S. J.P. developed the theoretical model. The manuscript was written by S.S. and T.H. All authors discussed the results and commented on the manuscript.

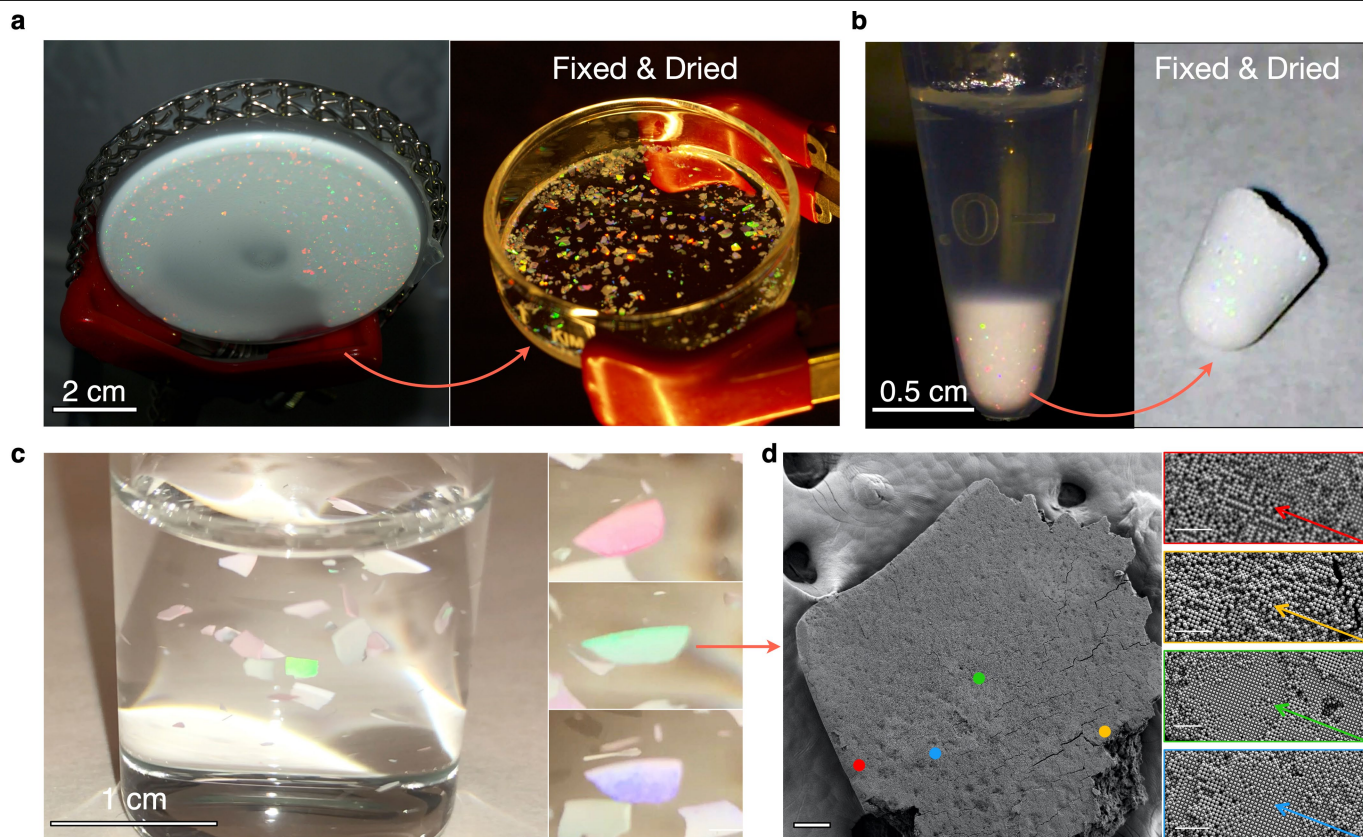
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2205-0>.

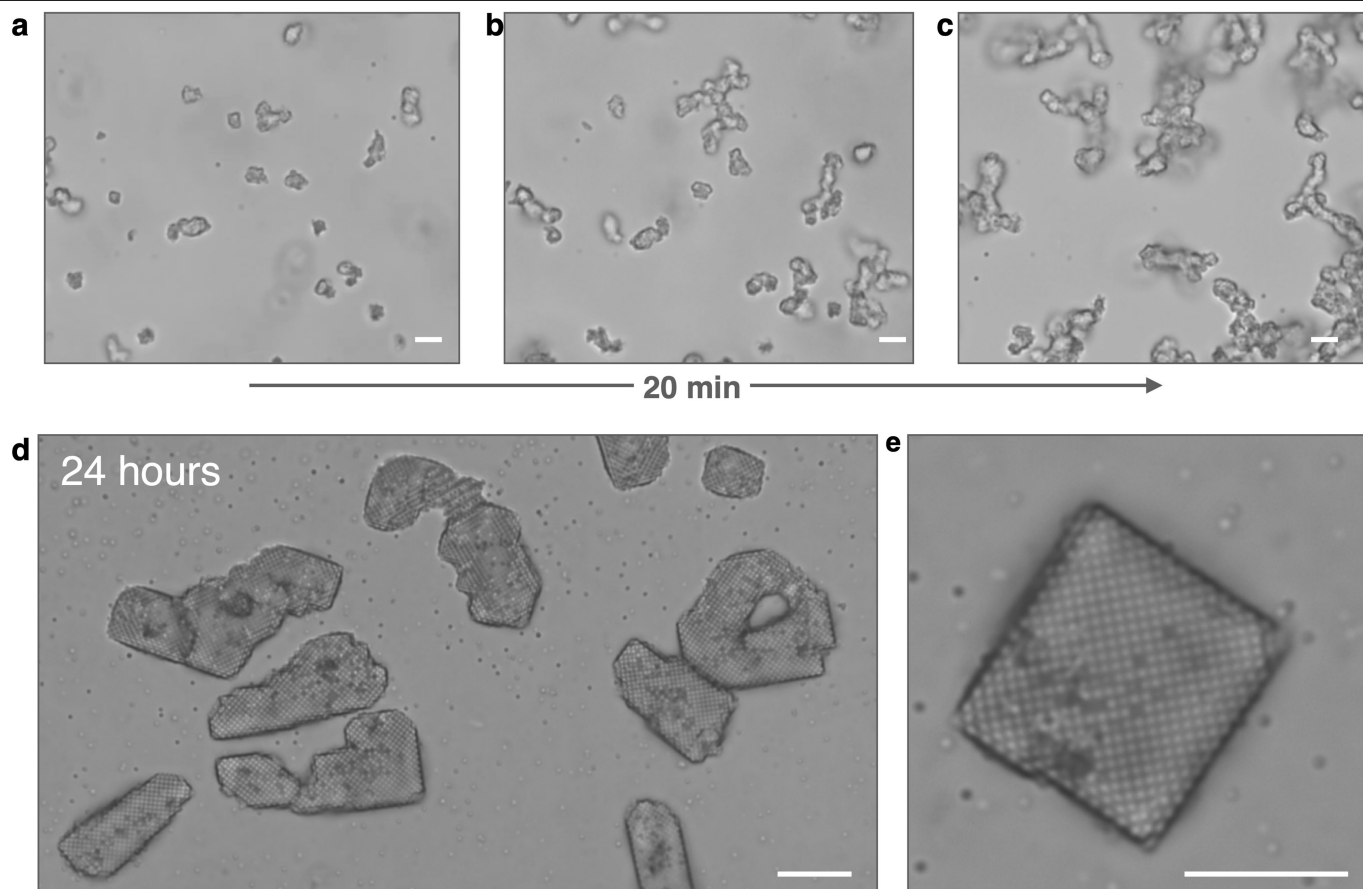
Correspondence and requests for materials should be addressed to S.S.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



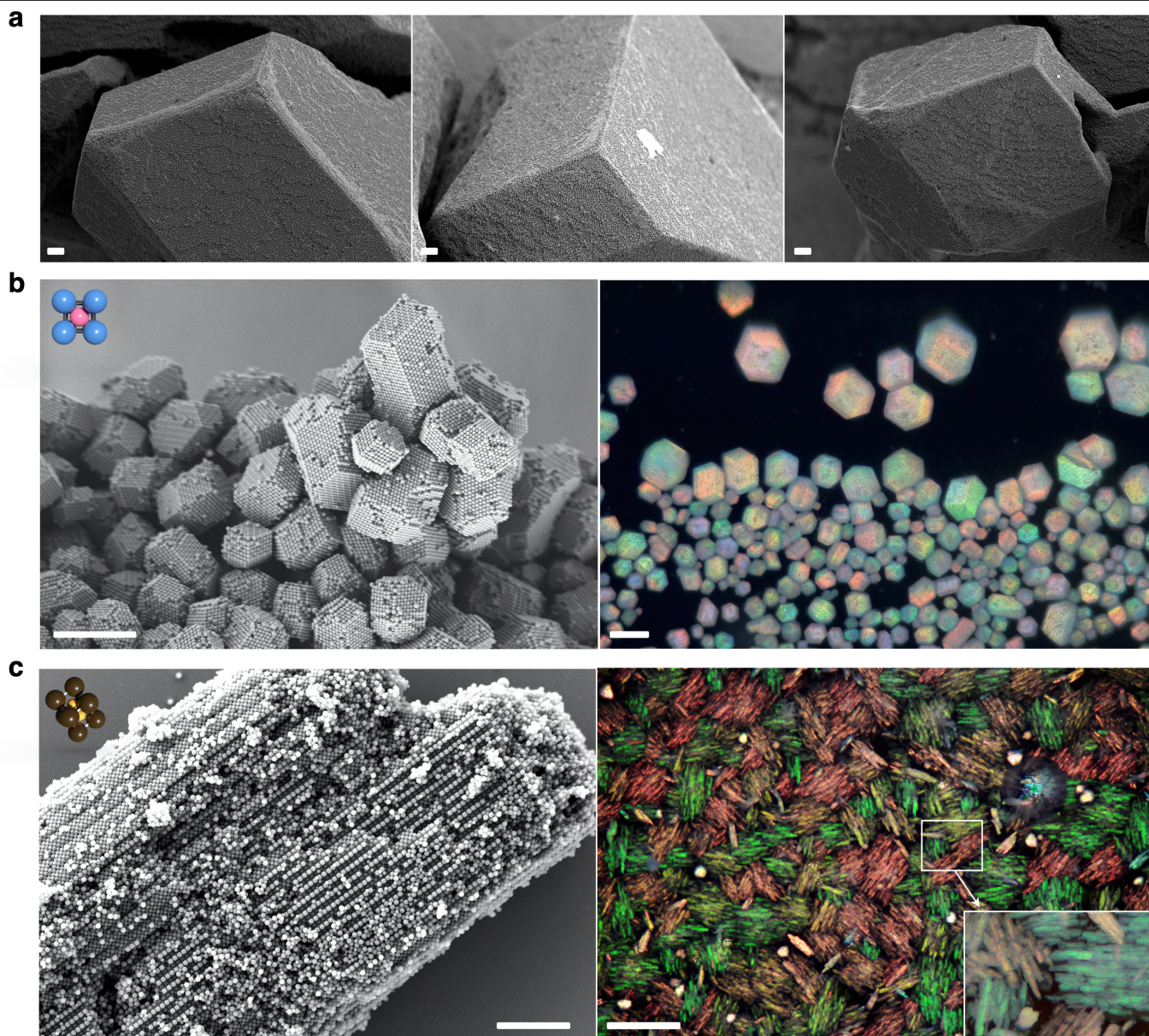
Extended Data Fig. 1 | Iridescent macroscopic single crystals. **a**, Single crystals with millimetre diameters visible on the bottom of a 10-cm Petri dish. Boundaries between single crystals are clear through the uniformly coloured regions of the Voronoi pattern that forms. **b**, The crystallinity of the bulk sediment is visible through iridescence when left undisturbed. At 5 mM of NaCl, the assembly is still reconfiguring and delicate. After fixing through dilution and slowly drying, the hardened sediment retains iridescence, denoting retention of crystalline order. In air versus water, iridescence is muted due to a

large refractive index mismatch. **c**, Fixed crystals dispersed and freely floating in density matched water (Supplementary Video 4). Slow rotation of the crystals reveals their colouration's angle dependence, displaying vibrant and uniformly coloured red, green, and blue iridescence as they rotate. **d**, SEM micrograph of a single crystal, examined at high magnification at four locations across the surface of the crystal. At each site, the crystal displays the same crystallographic plane and angle, demonstrating that the whole crystal is in register. Scale bar, 100 μm . Scale bar of high magnification sites, 5 μm .



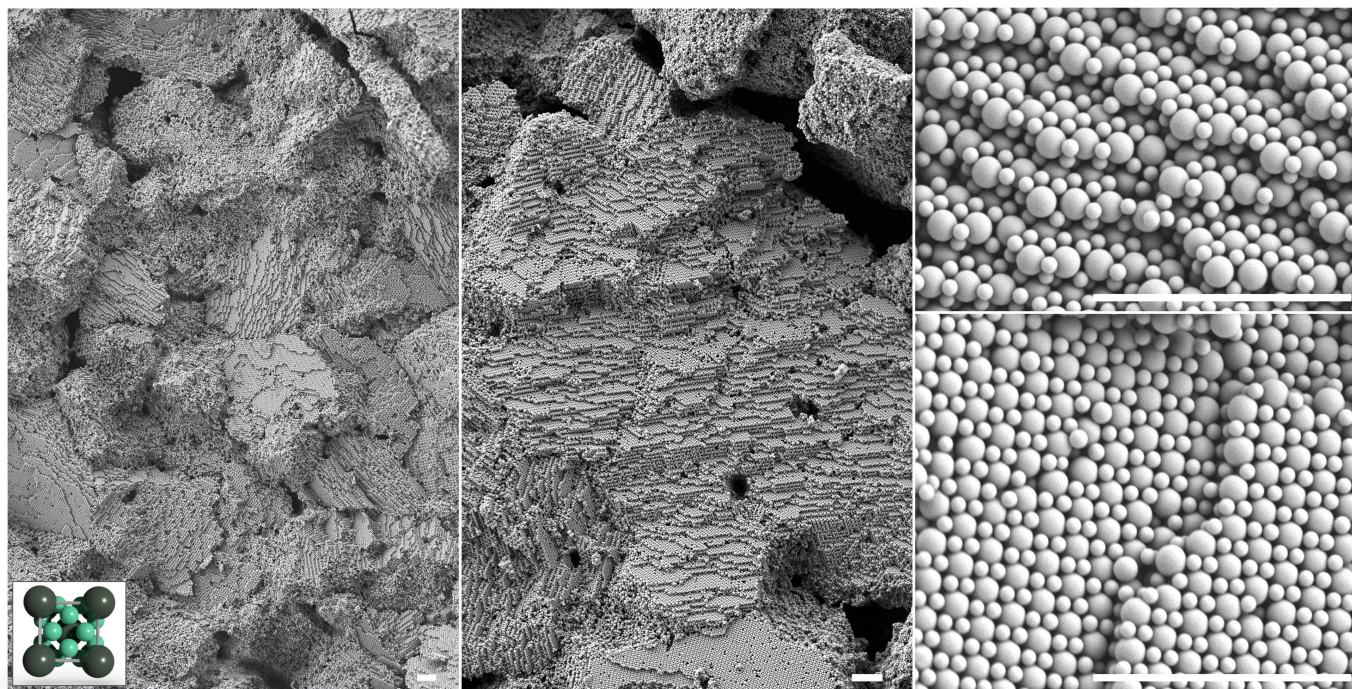
Extended Data Fig. 2 | Crystals annealing with strong electrostatic attraction. **a–c**, A 20-min time lapse of particles calculated to have a potential well depth of 12 kT. Coagulation occurs immediately after mixing and disordered heteroaggregates rapidly sediment. Although no order is present, a small degree of reconfiguration can be observed on the single-particle level.

Images were taken at 1 min (**a**), 5 min (**b**) and 20 min (**c**). **d**, After 24 h, the sample has crystallized. **e**, Crystals produced in this manner display the (100) plane due to the strong electrostatic attraction to the negatively charged substrate. Scale bars, 10 μm .

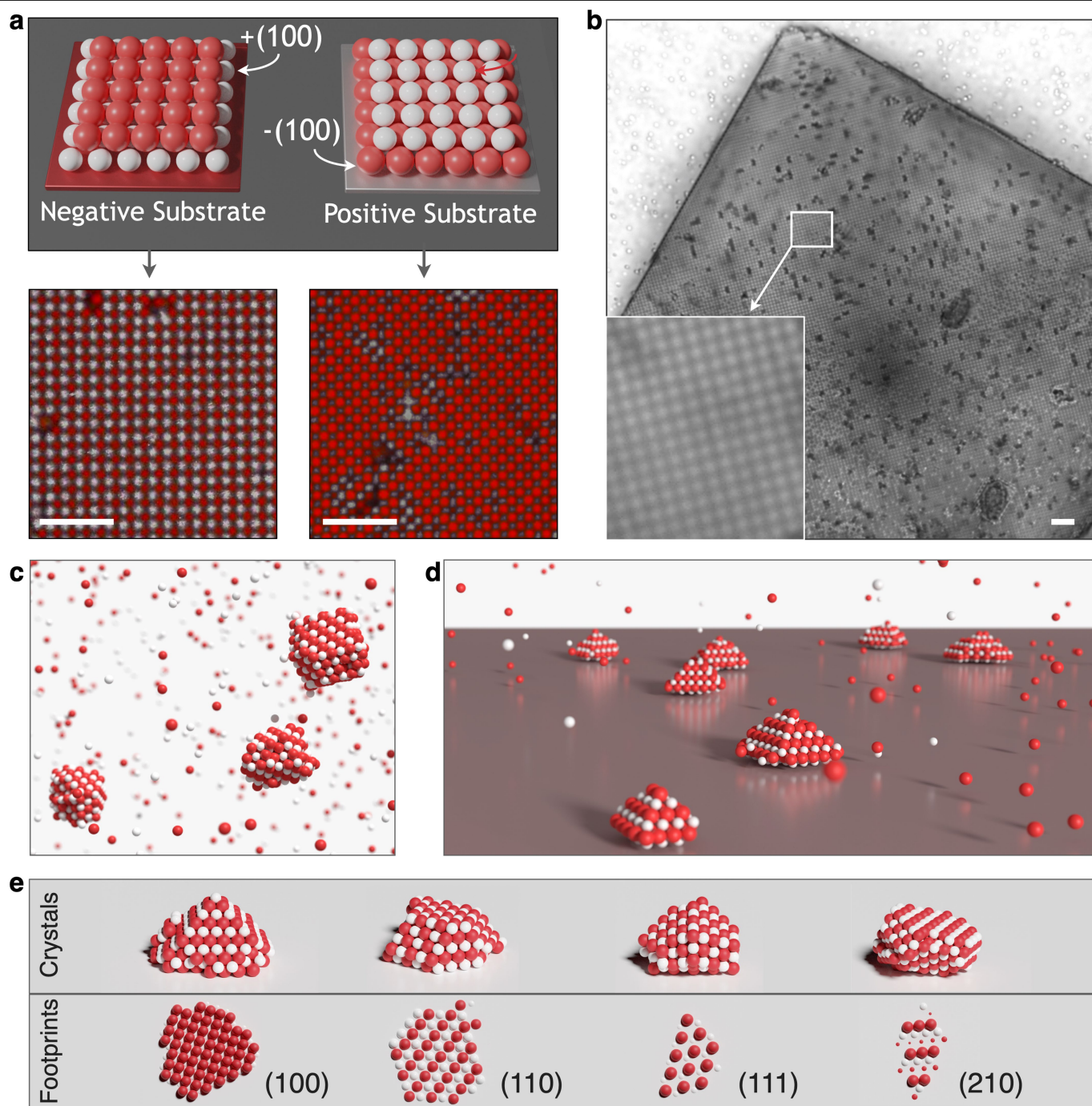


Extended Data Fig. 3 | Crystal habits. **a**, SEM micrographs of crystal facets from a rhombic dodecahedral CsCl ionic colloidal crystal. Scale bars, 10 μm . **b**, Left, SEM micrograph of small CsCl crystals with dodecahedral habit. Close inspection reveals sides comprised of the (110) plane. Scale bar, 10 μm . Right, optical micrograph of CsCl crystals. Scale bar, 100 μm . **c**, Left, SEM micrograph

of an AlB_2 needle. Close inspection reveals crystals growing with a railroad pattern characteristic of their (1010) plane. Scale bar, 3 μm . Right, optical micrograph of needles growing aligned in bunches and reveal the same colour, in sets of red and greens depending on the angle of orientation. Scale bar, 120 μm .



Extended Data Fig. 4 | Layered pattern of K_4C_{60} crystals. SEM micrographs of a bulk K_4C_{60} ionic colloidal crystal captured at increasing magnification. Cleavage of the crystals reveals sheets of the (110) plane. Scale bars, 5 μm .



Extended Data Fig. 5 | Substrate charge influence on heterogeneous nucleation of crystals. **a**, Schematic of substrate charge influence, where the (100) plane of either charge grows on an oppositely charged substrate. Bottom left, (100) plane of positive (white) particles growing on a naturally negatively charged glass substrate. Bottom right, glass surface treated with aminosilane for positive charge nucleates a negative (100) plane. Scale bars, 6 μm . **b**, Bright field microscopy image showing a macroscopic crystal nucleated on a negatively charged substrate. The magnified inset reveals the characteristic

pattern of the (100) plane in contact with the substrate. **c**, Image of faceted CsCl crystals that nucleate spontaneously in simulation, where $\lambda_D = 4.5 \text{ nm}$, $r_N = 500 \text{ nm}$ and $r_P = 430 \text{ nm}$. **d**, Heterogeneous nucleation is observed when a model attractive surface is added under the same conditions as **b**. The attractive surface has a strength of $\epsilon = 5k_B T$ for positive particles. **e**, Structures heterogeneously assembled relative to these facets were observed with an attractive and repulsive strength of $\epsilon = 3k_B T$ and with an increased size of N particles such that the ratio of $\sigma_N/\sigma_P = 1.24$.

Extreme rainfall triggered the 2018 rift eruption at Kīlauea Volcano

<https://doi.org/10.1038/s41586-020-2172-5>

Jamie I. Farquharson^{1✉} & Falk Amelung¹

Received: 24 July 2019

Accepted: 11 February 2020

Published online: 22 April 2020

 Check for updates

The May 2018 rift intrusion and eruption of Kīlauea Volcano, Hawai'i, represented one of its most extraordinary eruptive sequences in at least 200 years, yet the trigger mechanism remains elusive¹. The event was preceded by several months of anomalously high precipitation. It has been proposed that rainfall can modulate shallow volcanic activity^{2,3}, but it remains unknown whether it can have impacts at the greater depths associated with magma transport. Here we show that immediately before and during the eruption, infiltration of rainfall into Kīlauea Volcano's subsurface increased pore pressure at depths of 1 to 3 kilometres by 0.1 to 1 kilopascals, to its highest pressure in almost 50 years. We propose that weakening and mechanical failure of the edifice was driven by changes in pore pressure within the rift zone, prompting opportunistic dyke intrusion and ultimately facilitating the eruption. A precipitation-induced eruption trigger is consistent with the lack of precursory summit inflation, showing that this intrusion—unlike others—was not caused by the forceful intrusion of new magma into the rift zone. Moreover, statistical analysis of historic eruption occurrence suggests that rainfall patterns contribute substantially to the timing and frequency of Kīlauea's eruptions and intrusions. Thus, volcanic activity can be modulated by extreme rainfall triggering edifice rock failure—a factor that should be considered when assessing volcanic hazards. Notably, the increasingly extreme weather patterns associated with ongoing anthropogenic climate change could increase the potential for rainfall-triggered volcanic phenomena worldwide.

Compelling evidence exists for seismicity generated by rainfall^{4,5}. Rainfall-induced stress changes at depth can promote fault initiation and reactivation—a mechanism invoked for numerous geological phenomena, including landslides⁶, silent slip events⁷ and remote triggering of earthquakes⁸. Rainfall can also interact with hot volcanic lava domes, causing gravitational dome collapse⁹, explosions² and the generation of lahars and other flow phenomena¹⁰. Explosive activity has even been linked to specific weather systems^{2,3}. These mechanisms probably influence volcanic activity only in the upper tens or hundreds of metres¹¹, prompting the suggestion that rainfall may only be a viable trigger for volcanic activity in the shallow subsurface^{12,13}. The only studies to link precipitation to deeper processes consider hydrological loading and unloading of the edifice^{14,15}. The question as to whether and how rainfall can directly induce deep magmatic activity remains unanswered.

In 2018, coincident with prolonged and extreme rainfall, Kīlauea underwent a complex, multistage eruption involving an extensive rift eruption and the collapse of the summit caldera. On 30 April a train of seismicity initiated along the East Rift Zone (ERZ) that was interpreted as a downrift dyke intrusion¹, ultimately breaking ground as a fissure eruption on 3 May, and followed the next day by a magnitude-6.9 earthquake. The summit exhibited numerous explosive eruptions and caldera collapse events that continued through to August; activity at the Lower ERZ was characterized by fissuring and lava fountaining.

This represents one of Kīlauea's most remarkable eruptive sequences over the past two centuries, not least because the initiation mechanism remains equivocal¹. We investigate several lines of evidence that suggest that anomalous rainfall weakened the edifice by instigating an impulsive pressure wave that propagated to depth and modified the local effective stress in the rift zone. We hypothesize that this in turn triggered the dyke intrusion and the eruptive aftermath.

Dyke initiation from a magma chamber can be considered in terms of the static tensile failure criterion around a cavity. Mechanical failure and dyke initiation will be achieved once the magma overpressure, δp_f , achieves a threshold value defined by¹⁶:

$$\delta p_f = \kappa(\rho g z - [p + \delta p] + \tau) \quad (1)$$

where the product of the rock density, ρ , gravitational acceleration, g , and depth, z , is the lithostatic stress; p and δp are the hydrostatic pore pressure and the pore-pressure change, respectively; and τ is the tensile strength of the host rock. In an elastic medium, the magma overpressure is proportional to the tangential stress at the chamber wall, and their ratio, κ , is a function of the chamber geometry: for example, $\kappa = 2$ for a spherical chamber within an infinite space¹⁷. Equation (1) highlights that—all other things held equal—an increase in pore pressure will decrease the failure overpressure required to initiate chamber-wall

¹Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, FL, USA. ✉e-mail: james.farquharson@rsmas.miami.edu

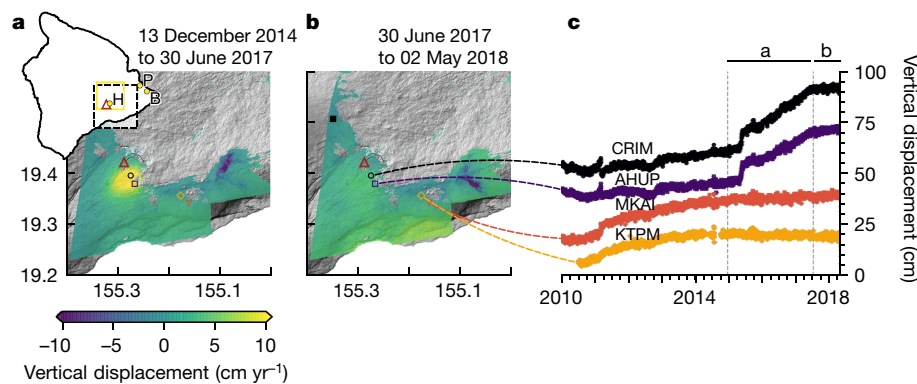


Fig. 1 | Pre-eruption ground deformation of the study site. a, Vertical deformation map derived from Sentinel-1 InSAR data (ascending track 124 and descending track 87) between December 2014 and June 2017. Inset, map showing the location of Kilauea Volcano (red triangle) and the HVNP rain gauge (H) within the Eastern Rift Zone. The yellow box outlines the $0.25^\circ \times 0.25^\circ$ TRMM/GPM footprint used here. The Paradise Park well (P) and Hawai'iian Beaches rain gauge (B) are also highlighted (see Extended Data Fig. 1). The

dashed box indicates the region shown in **a**, **b**. In **b**, the filled square is a reference point; in **a**, **b**, open symbols represent GPS station locations. **b**, Vertical deformation between June 2017 and May 2018. **c**, Vertical GPS displacement of the Kilauea flank from 2010. Station names are shown on the panel; colours correspond to locations in **a**, **b**. Dashed vertical lines highlight marked changes-of-slope in station data proximal to the summit (for CRIM and AHUP). Datasets have been offset from the x axis for clarity.

failure and dyke initiation by the amount $\kappa\delta p$. Increasing fluid pressure facilitates not only the initiation of a dyke but also its propagation, in the sense that the fracture toughness (or energy) of the edifice material is dependent on the ambient stress field¹⁸ and thus on the pore pressure. Effective stress (that is, lithostatic stress minus pore pressure) governs the failure stress of volcanic materials¹⁹. During and after heavy rainfall, an impulsive increase in groundwater volume will cause a perturbation in pore pressure at depth, increasing it transiently above the background hydrostatic condition. Under unconfined, saturated conditions typical of basaltic systems, such pore-pressure transients are often reflected by water-level changes in nearby wells²⁰; this is demonstrated in Extended Data Fig. 1, where we correlate rainfall with recorded well data from a site 15 km north of the rift zone. In theory—if the volume and rate of meteoric water infiltration are sufficient—mechanical failure can be induced in the vicinity of a magma chamber or dyke, in turn triggering intrusion or eruption.

Interferometric synthetic-aperture radar (InSAR) data show that more than 0.3 m of uplift occurred at the summit between 2014 and mid-2017 (Fig. 1a), yet only around 0.01 m of uplift was detected at the summit over the following 10 months (Fig. 1b). This temporal deformation pattern is confirmed by data from GPS stations proximal to Kilauea's caldera (CRIM and AHUP), showing that moderate inflation occurred between 2010 and 2015, followed by a substantial increase in inflation rate until around June 2017 (Fig. 1c), after which inflation was negligible for the following 10 months. In the upper and middle ERZ, no substantial inflation was detected between 2010 and 2018 (Fig. 1b; GPS stations MKAI and KTPM in Fig. 1c). This implies that the inflation observed from mid-March¹ reflects local, shallow processes, rather than wholesale pressurization of the system. The lack of precursory summit and rift-zone inflation suggests that the intrusion–eruption was not triggered by an influx of fresh magma from depth but that it was a passive intrusion, caused by extension and/or weakening of the rift zone^{21,22}.

In early 2018, the Hawai'iian islands were subject to protracted, at times extreme, rainfall (Fig. 2). The maximum peak in the rainfall power spectrum occurs at 1 yr^{-1} , indicating substantial annual seasonality in rainfall over Kilauea Volcano on Hawai'i Island (generally, most rainfall occurs between 9 March and 25 August), overlain by a non-negligible stochastic component (Fig. 2b). This annual signal accounts for over half of the variability in rainfall over Kilauea; however, an aseasonal shift of the synoptic-scale atmospheric wave pattern across the North Pacific in mid-to-late January 2018 preceded the passage of several consecutive low-pressure systems over Hawai'i in the months to follow.

Several months of greater than average rainfall culminated in record downpour, with 1.26 m of rain falling within 24 h (14–15 April 2018) on Kauai Island (northwest of Hawai'i Island)—a record not only for Hawai'i but for the entire United States²³. Calibrated precipitation data from NASA's Tropical Rainfall Measurement Mission and Global Precipitation Measurement mission (TRMM/GPM) satellites indicate that over 2.25 m of rainfall fell over Kilauea during the first quarter of 2018, compared with a first-quarter 19-year average of 0.90 m. Figure 2c, e shows the multiday cumulative sum of rainfall calculated as a moving window across the time series, using windows of 30 days (Fig. 2c, approximately 1 month) and 180 days (Fig. 2e, approximately 6 months). These data are lognormal (Fig. 2d, f). Notably, the 30-day total rainfall exceeds 2 standard deviations of the mean ($+2\sigma$) immediately before the 2018 flank eruption: that is, a statistically significant deviation. Even more strikingly, the rolling 180-day cumulative rainfall has only two periods in which rainfall exceeds this threshold, one of which directly precedes the eruption.

Kilauea Volcano is hydrogeologically complex²⁴, as highlighted by laboratory data (for example, see ref. ²⁵; see Methods). Accordingly, we model the edifice as two connected saturated domains in a one-dimensional half-space (model α): a highly permeable shallow layer (0–500 m) overlying an intermediate-permeability domain (0.5–10 km). Propagation of pore pressure p from the surface owing to precipitation is modelled using a finite difference approximation to solve for transient, vertical flow of groundwater (that is, the diffusion problem), such that $(\partial p / \partial t) = D(\partial^2 p / \partial z^2)$, where z is depth, t is time and D is hydraulic diffusivity (a function of permeability, bulk modulus, fluid viscosity and porosity). Full details, including parameter values, are given in Methods. For model α , $D = 37 \text{ m}^2 \text{ s}^{-1}$ and $0.34 \text{ m}^2 \text{ s}^{-1}$ above and below 0.5 km (below the surface), respectively (Fig. 3a). Using the calibrated satellite data as a variable-flux boundary, we show the resultant maximum pore-pressure change in Fig. 3b. Three additional models are shown for reference to demonstrate the range of feasible pressure responses to rainfall (see Methods and Extended Data Table 1).

In all modelled scenarios, we observe a quasistatic pore-pressure build-up of tens to thousands of pascals at depths 1–6 km below the surface immediately before the onset of the 2018 flank eruption. For completeness, we include results from the unlikely end-member scenarios in Extended Data Fig. 2. More complex and realistic models yield intermediate pressure changes (Fig. 3b and Extended Data Fig. 2), as demonstrated in Fig. 4a–c. The April 2018 peak in pore pressure is the highest observed throughout the modelled period (for example, see Fig. 4b, c). Parametric analysis of each of the model frameworks shown

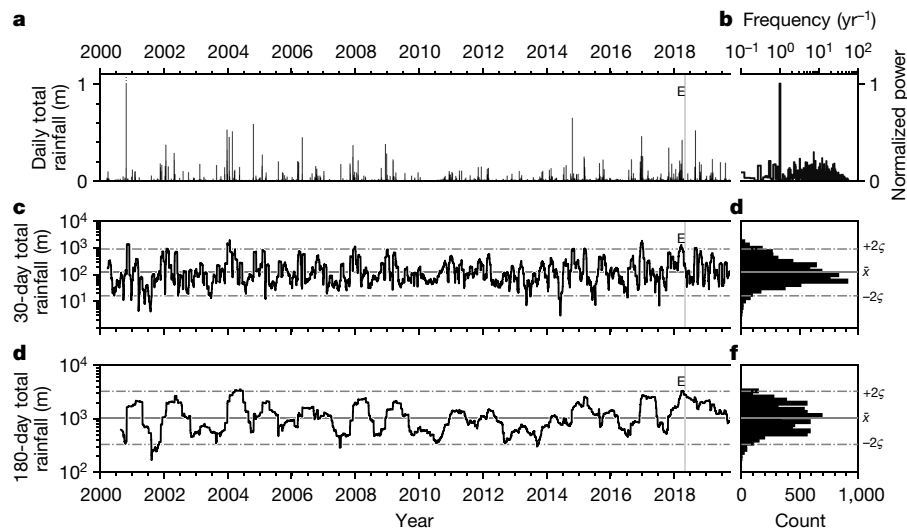


Fig. 2 | Rainfall over Kilauea. **a**, Calibrated daily rainfall amount over Kilauea from March 2000 to July 2019 (TRMM/GPM data calibrated with gauge data; see Fig. 1a inset and Methods). The vertical bar shows the date of flank eruption (E). **b**, Power spectrum of Fourier-transformed rainfall time series (a) shown in the (normalized) power–frequency domain. **c**, Rolling 30-day

cumulative rainfall since March 2000 at Kilauea. Horizontal dashed lines delineate two standard deviations either side of the mean ($\bar{x} \pm \sigma$). **d**, Histogram of data in **c** on log-linear axes. **e**, Rolling 180-day cumulative rainfall. **f**, Histogram of data from **e** on log-linear axes.

in Fig. 3a reveals that this pattern holds for a wide range of feasible physical properties. In our preferred model (model α), we obtain pressure changes of around 0.1 kPa and 1 kPa at 3 km and 1 km below the surface, respectively.

At depths of around 3 km below the surface (or 1.8 km below sea level (b.s.l.), the depth estimated for most lateral magma transport in the rift zone¹), the elevated subsurface pore pressure in early 2018 was the

largest peak in pressure in 47 years—the highest since the onset of the Pu’u ‘Ō’ō eruption (Fig. 4b, c). Given the mechanism described above, this is a strong indicator that elevated pore fluid pressures facilitated the 2018 intrusion and rift eruption. Moreover, we outline four independent lines of evidence to support this proposition.

First, there was very little precursory inflation immediately before the rift eruption (Fig. 1b, c). The rapid uplift recorded by tiltmeters in the Pu’u ‘Ō’ō area from mid-March reported in ref. ¹ thus reflects a highly localized, shallow deformation source. The absence of widespread precursory inflation suggests that the 2018 intrusion was passive, fostered by weakening of the rift zone.

Second, a statistical analysis of the Kilauea’s reported historical eruptions shows that the volcano exhibits a marked tendency towards erupting during the wettest times of the year (Extended Data Fig. 3): the onset of around 60% of reported eruptions since 1790 (including the Pu’u ‘Ō’ō eruption 1983–2018) occurred during the ‘rainy’ season, despite the fact that Kilauea’s ‘rainy’ season is shorter than its ‘dry’ season (see Methods).

Third, recorded intrusions appear to be correlated with elevated pore pressures at depth. Figure 4b shows intrusions into the rift zone since the 1975 Kalapana earthquake (compiled after refs. ^{22,26–28}). Comparing the modelled rainfall-induced pressure perturbation at a depth of 3 km below the surface with the long-term average, more than 60% of intrusions (20 of 33) are associated with periods of above-average pore pressure. If we compare the pore-pressure change with the rolling four-year average at the same depth in order to account for interannual fluctuations (such as the El Niño–Southern Oscillation) we find that 19 (58%) were initiated when the pore pressure was above this threshold. Clearly, not every peak in pore pressure is associated with an intrusion, and vice versa—a function of superposing processes with different periodicities (a combination of internal and external forcing), and thus characteristic of triggered systems (see, for example, ref. ¹⁴). Nevertheless, it is striking that intrusions are approximately twice as likely to occur at Kilauea when pore pressure is elevated, suggesting that pore pressure can act as a trigger mechanism in a critically stressed—or ‘primed’—volcanic system.

Finally, historical precipitation records show that Kilauea’s May 1924 eruption—the previous extraordinary eruption—also followed extremely wet conditions. In April 1924, many stations across Hawai’i recorded as much as 0.5 m in excess of the long-term average for that

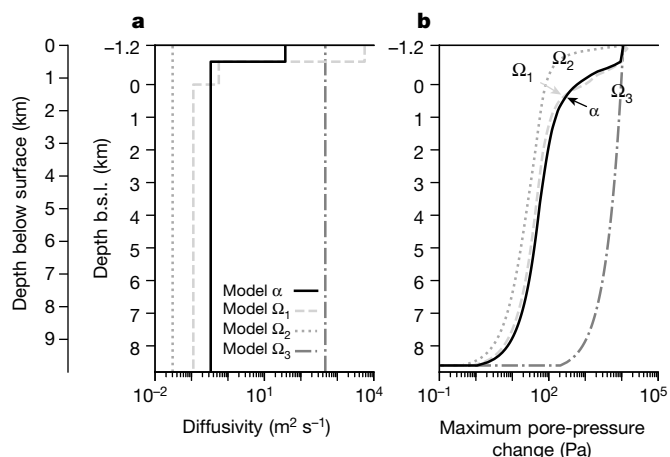


Fig. 3 | Diffusion model metadata. **a**, Diffusivity (a function of permeability, k , bulk modulus, K , fluid viscosity, μ , and porosity, ϕ , given by $kK/\mu\phi$) against depth (depth below the surface and depth b.s.l. are shown for reference) for four different models ($k-\phi$ scenarios, summarized in Extended Data Table 1). The primary model (model α) is a two-layer model, with a zone of high porosity and permeability overlying a lower permeability domain. Reference model Ω_1 subdivides the lower domain into two, each with different properties. Models Ω_2 and Ω_3 are single-domain models, meaning that they use values of porosity, permeability, bulk modulus and viscosity that are constant with depth. See Methods for more details. **b**, The maximum pressure change effected at depth in each model throughout the modelled time series. Note that the maximum pressure change for any given depth can occur on numerous dates and may not be consecutive. For example, high values near the surface may occur frequently (see Fig. 4a, c) but may not propagate to depth. On the other hand, high pressure at depth may not be a direct consequence of a single spike at the surface.

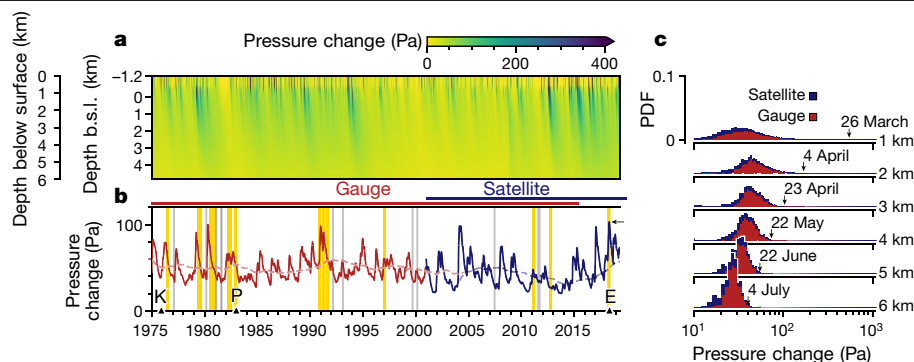


Fig. 4 | Pore-pressure change in response to infiltration into Kilauea's edifice. a, Pore-pressure change modelled over the period January 1975 to April 2019, using available HVNP gauge data (1950–2015) and calibrated satellite data (2000–2019), with depth and time. Daily rainfall data are used as a fluctuating boundary condition. Colour scale indicates pressure change. **b**, Pore-pressure change at 3 km below the surface (1.8 km b.s.l.) modelled over the period January 1950 to April 2019 (data shown are since the 1975 Kalapana earthquake). The dashed line shows the four-year running average. K represents the 1975 magnitude-7.2 Kalapana earthquake; P shows the 1983 onset of the Pu'u 'Ō'ō eruption; E represents the 2018 Kilauea rift intrusion–eruption. Vertical bars show reported intrusion events within the rift zone, after refs. ^{22,26–28}. Intrusions

month, and the gauge measurements for several of these were at that time the highest daily rainfall amounts on record²⁹. In particular, more than 0.1 m fell at the Hawai'i Volcano Observatory in 24 h (14 April). The 2018 eruption echoed many features of 1924 (for example, major summit explosions and a drop in lava lake level), suggesting that not only the timing of intrusions and eruptions but also the eruptive style of Kilauea is influenced by rainfall.

Diffusion modelling shows that rainfall can induce quasistatic pressure changes on the order of kilopascals to tens of kilopascals at depths of a few kilometres (Fig. 4 and Extended Data Fig. 2, respectively). Increasing pore pressure can cause embrittlement³⁰, hydrofracture³¹ and dyke initiation³² by reducing the static threshold for tensile failure. Although the precise magnitude of stress change will vary owing to drainage and the thermal contribution of the magma, pressure changes on the order of 1 kPa are in line with trigger stresses caused by solid Earth tides^{33–35} and those required to trigger earthquakes on pre-existing faults³⁶. Pressure changes on the order of 10 kPa are typically assumed to be necessary to trigger mechanical failure and attendant geological processes in unstressed media⁴; however, it has been demonstrated that stress changes in the range of 0.1–1 kPa are sufficient for rainfall to trigger earthquakes provided that the crust is in a critical state⁵. It has been inferred from historical eruption patterns that Kilauea is particularly sensitive to external modulation and eruption triggering²⁶; moreover, recent evidence shows that the rift zone has undergone mechanical weakening over the course of the Pu'u 'Ō'ō eruption³⁷. Although we do not account here for potential amplification or dampening of the precipitation-induced pressure change in response to poroelastic effects, our results show that the quasistatic stress changes associated with pore-pressure infiltration can effect a sustained increase above long-term background levels in a way that dynamic stresses of similar magnitude (for example, solid Earth tides) cannot. Relatively small changes in edifice pore pressure can result in a substantial reduction in the overpressure required to instigate magma chamber failure (Equation (1)). As such, precipitation-induced pore-pressure fluctuations contribute to the overall stress state of Kilauea Volcano; we propose that this hydromechanical coupling may directly trigger primary volcanic activity.

The unprecedented rainfall over Hawai'i in the months before the 2018 flank eruption increased the potential for mechanical failure within the edifice. Taken together, the separate lines of evidence

are highlighted in yellow if they coincide with periods during which pressure change exceeds the four-year average, and grey if they do not. Intrusion 33 in this time series corresponds to the early 2018 activity (intrusion detected mid-March, followed by the rift eruption on 3 May). The arrow highlights the maximum pore-pressure perturbation over this timeframe (1975 to 2019), coinciding with the onset of 2018's rift eruption. Horizontal bars indicate data availability. **c**, Probability density function (PDF) of modelled pressure change at depths 1–6 km below the surface. Arrows highlight the pore-pressure front diffusing from near the surface (1 km) to greater depths over time (26 March, 4 April, 23 April, 22 May, 22 June and 4 July 2018 at 1 km, 2 km, 3 km, 4 km, 5 km and 6 km below surface).

reported above strongly suggest a correlation between rainfall and volcanic activity at Kilauea—not only in 2018, but throughout its eruptive history. By locally reducing effective stress at depth, prolonged periods of rainfall may induce opportunistic dyke intrusions or facilitate dyke propagation. The historical preponderance of Kilauea's eruptions during the wettest parts of the year buttresses this theory, as does the coincidence of dyke intrusions with elevated subsurface pore pressure and the similarities observed between volcanic events associated with similar rainfall patterns. Critically, as our climate continues to change, the occurrence of prolonged periods of extreme rainfall is predicted to increase in many parts of the world, increasing the potential for rainfall-triggered volcanic phenomena. Elevated pore pressures at depth tend to be fostered and maintained by prolonged periods of above-average rainfall, associated with long-lived and generally forecastable synoptic-scale systems: by better understanding the hydromechanical couplings between rainfall and volcanism, advanced warning of rainfall-induced volcanic hazards may be achievable.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2172-5>.

1. Neal, C. A. et al. The 2018 rift eruption and summit collapse of Kilauea Volcano. *Science* **363**, 367–374 (2019).
2. Mastin, L. G. Explosive tephra emissions at Mount St. Helens, 1989–1991: the violent escape of magmatic gas following storms? *Geol. Soc. Am. Bull.* **106**, 175–185 (1994).
3. Matthews, A. J. & Barclay, J. A thermodynamical model for rainfall-triggered volcanic dome collapse. *Geophys. Res. Lett.* **31**, L05614 (2004).
4. Costain, J. K. Groundwater recharge as the trigger of naturally occurring intraplate earthquakes. *Geol. Soc. Lond. Spec. Publ.* **432**, 91–118 (2017).
5. Hainzl, S., Kraft, T., Wassermann, J., Igel, H. & Schmedes, E. Evidence for rainfall-triggered earthquake activity. *Geophys. Res. Lett.* **33**, L19303 (2006).
6. Handwerker, A. L., Rempel, A. W., Skarbek, R. M., Roering, J. J. & Hilley, G. E. Rate-weakening friction characterizes both slow sliding and catastrophic failure of landslides. *Proc. Natl Acad. Sci. USA* **113**, 10281–10286 (2016).
7. Kodaira, S. et al. High pore fluid pressure may cause silent slip in the Nankai Trough. *Science* **304**, 1295–1298 (2004).
8. Prejean, S. G. et al. Remotely triggered seismicity on the United States west coast following the M_w 7.9 Denali Fault earthquake. *Bull. Seismol. Soc. Am.* **94**, S348–S359 (2004).

9. Voight, B., Constantine, E. K., Siswoidjono, S. & Torley, R. Historical eruptions of Merapi volcano, central Java, Indonesia, 1768–1998. *J. Volcanol. Geotherm. Res.* **100**, 69–138 (2000).
10. Yamasato, H., Kitagawa, S. & Komiya, M. Effect of rainfall on dacitic lava dome collapse at Unzen volcano, Japan. *Pap. Meteorol. Geophys.* **48**, 73–78 (1998).
11. Christenson, B. W. et al. Cyclic processes and factors leading to phreatic eruption events: insights from the 25 September 2007 eruption through Ruapehu Crater Lake, New Zealand. *J. Volcanol. Geotherm. Res.* **191**, 15–32 (2010).
12. Canon-Tapia, E. Volcanic eruption triggers: a hierarchical classification. *Earth Sci. Rev.* **129**, 100–119 (2014).
13. Matthews, A. J., Barclay, J. & Johnstone, J. E. The fast response of volcano-seismic activity to intense precipitation: triggering of primary volcanic activity by rainfall at Soufrière Hills Volcano, Montserrat. *J. Volcanol. Geotherm. Res.* **184**, 405–415 (2009).
14. Violette, S. et al. Can rainfall trigger volcanic eruptions? A mechanical stress model of an active volcano: ‘Piton de la Fournaise’, Reunion Island. *Terra Nova* **13**, 18–24 (2001).
15. Hammond, W. C., Kreemer, C., Zaliapin, I. & Blewitt, G. Drought-triggered magmatic inflation, crustal strain and seismicity near the Long Valley Caldera, Central Walker Lane. *J. Geophys. Res. Solid Earth* **124**, 6072–6091 (2019).
16. Timoshenko, S. P. & Goodier, J. *Theory of Elasticity* 2nd edn (Maple Press, 1951).
17. Tait, S., Jaupart, C. & Vergnolle, S. Pressure, gas content and eruption periodicity of a shallow, crystallising magma chamber. *Earth Planet. Sci. Lett.* **92**, 107–123 (1989).
18. Rubin, A. M. Tensile fracture of rock at high confining pressure: implications for dike propagation. *J. Geophys. Res. Solid Earth* **98** (B9), 15919–15935 (1993).
19. Wong, T. F. & Baud, P. The brittle-ductile transition in porous rock: a review. *J. Struct. Geol.* **44**, 25–53 (2012).
20. Jónsson, S., Segall, P., Pedersen, R. & Björnsson, G. Post-earthquake ground movements correlated to pore-pressure transients. *Nature* **424**, 179–183 (2003).
21. Owen, S. et al. January 30, 1997 eruptive event on Kilauea Volcano, Hawaii, as monitored by continuous GPS. *Geophys. Res. Lett.* **27**, 2757–2760 (2000).
22. Poland, M. P., Takahashi, T. J. & Landowski, C. M. *Characteristics of Hawaiian Volcanoes* Professional Paper 1801 (Government Printing Office, 2014).
23. National Climate Extremes Committee. *National Record 24-Hour Precipitation at Waipā Garden, Hawai‘i* Report 14 December 2018 (National Centers for Environmental Information, 2018).
24. Ingebritsen, S. E. & Scholl, M. A. The hydrogeology of Kilauea volcano. *Geothermics* **22**, 255–270 (1993).
25. Keller, G. V., Grose, L. T., Murray, J. C. & Skokan, C. K. Results of an experimental drill hole at the summit of Kilauea Volcano, Hawaii. *J. Volcanol. Geotherm. Res.* **5**, 345–385 (1979).
26. Klein, F. W., Koyanagi, R. Y., Nakata, J. S. & Tanigawa, W. R. in *U.S. Geological Survey Professional Paper, Issue 1350* Vol. 2 (eds Decker, R. W. et al.) 1019–1185 (US Geological Survey, 1987).
27. Heliker, C. & Mattox, T. N. *The First Two Decades of the Pu‘u ‘Ō‘ō-Kūpaianaha Eruption: Chronology and Selected Bibliography* (U.S. Geological Survey, 2003).
28. Orr, T. R. et al. in *Hawaiian Volcanoes: From Source to Surface* Vol. 208, Ch. 19 (eds Carey, R. et al.) 393–420 (2015).
29. Loveridge, E. F. April 1924 climatological data: Hawaii section. *Climatological Data* **4**, 25–40 (1924).
30. Farquharson, J., Heap, M. J., Baud, P., Reuschlé, T. & Varley, N. R. Pore pressure embrittlement in a volcanic edifice. *Bull. Volcanol.* **78**, 6 (2016).
31. Gudmundsson, A. *Rock Fractures in Geological Processes* (Cambridge Univ. Press, 2011).
32. Albino, F., Amelung, F. & Gregg, P. The role of pore fluid pressure on the failure of magma reservoirs: insights from Indonesian and Aleutian arc volcanoes. *J. Geophys. Res. Solid Earth* **123**, 1328–1349 (2018).
33. Tolstoy, M., Vernon, F. L., Orcutt, J. A. & Wyatt, F. K. Breathing of the seafloor: tidal correlations of seismicity at Axial volcano. *Geology* **30**, 503–506 (2002).
34. Scholz, C. H., Tan, Y. J. & Albino, F. The mechanism of tidal triggering of earthquakes at mid-ocean ridges. *Nat. Commun.* **10**, 2526 (2019).
35. Dzurisin, D. Influence of fortnightly earth tides at Kilauea volcano, Hawaii. *Geophys. Res. Lett.* **7**, 925–928 (1980).
36. Saar, M. O. & Manga, M. Seismicity induced by seasonal groundwater recharge at Mt. Hood, Oregon. *Earth Planet. Sci. Lett.* **214**, 605–618 (2003).
37. Wauthier, C., Roman, D. C. & Poland, M. P. Modulation of seismic activity in Kilauea’s upper East Rift Zone (Hawai‘i) by summit pressurization. *Geology* **47**, 820–824 (2019).

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Rainfall data

To obtain a contiguous rainfall time series at Kilauea, we extract precipitation data for the period March 2000 to June 2019 from the NASA/JAXA Tropical Rainfall Measuring Mission (TRMM) and Global Precipitation Measurement (GPM) satellite datasets, which are operationally available from NASA from March 2000. We use the 3B42 Research Version product (<https://doi.org/10.5067/TRMM/TMPA/3H/7>), a $0.25^\circ \times 0.25^\circ$ gridded daily product resampled from the original three-hourly rainfall estimates. Although this provides a daily rainfall estimate over the entire time frame, these data are an areal average and under-represent the true volume of rainfall at Kilauea by a factor, Λ , of 5.12. We calibrate the satellite data using rain gauge data available between 1950 and 2015 (gauge H in Fig. 1a, inset). The gauge data are from the Hawai'i Volcano National Park rainfall gauge (Global Historical Climatological Network Daily (GHCND) identification code USC00511303), available through the Climate Data Online project of the National Oceanic and Atmospheric Administration (NOAA). This gauge site was a volunteer observer station which closed as of 22 June 2015.

Ground deformation

Vertical velocity data were obtained by processing Sentinel-1 ascending and descending SAR data. Data were processed using the US Jet Propulsion Laboratory (JPL)-developed InSAR Scientific Computing Environment (ISCE) open-source software package³⁸, and further time-series analysis was performed using the MintPy software toolbox (Miami INsar Time-series software in PYTHON), developed at the University of Miami³⁹. Vertical GPS data (Fig. 1c) were accessed from the Nevada Geodetic Laboratory for GPS stations CRIM, AHUP, KTPM and MKAI (Fig. 1a, b; <http://geodesy.unr.edu/NGLStationPages/stations/>; ref. ⁴⁰). Vertical data are extracted directly with associated vertical error.

Modelling approach

We model the subsurface pore-pressure perturbation as a function of rainfall by using a finite difference approximation to solve for a transient, vertical flow of groundwater (that is, the diffusion problem), such that $(\partial p / \partial t) = [kK / \mu \phi] [\partial^2 p / \partial z^2]$. Here, pore pressure, p , over time, t , and depth, z , beneath the surface is a function of the permeability, k , the bulk modulus, K , and the porosity, ϕ , of the edifice, as well as the viscosity, μ , of the percolating fluid (water). The expression $kK / \mu \phi$ represents the (hydraulic) diffusion coefficient. Assuming a zero-flux boundary, we impose a pressure change of zero at the base, Z , of the domain, which is arbitrarily given as 10 km below the surface (8.8 km b.s.l.), that is, $(\partial p / \partial z)|_{z=Z} \approx \delta p|_{z=Z} = 0$. We note that this is deep enough that we do not observe boundary effects in the depth range of interest. Although subsurface pressure data for calibration are scarce, this solution appears to reflect pressure changes at depth well (Extended Data Fig. 1), as it relies on physical parameters.

Pressure at the surface is defined given the calibrated height of recorded rainfall, h , the density of water, ρ_w , and the acceleration due to gravity, g , such that $p(z=0, t) = \Lambda \rho_w g h(t)$, where $\Lambda = 5.12$ is the calibration factor described above, and z represents the depth below the ground surface. To test the sensitivity of our model to the rainfall input, we also ran it assuming daily rainfall values of $\pm 10\%$ of the recorded value: we note that the relative timing and magnitude of pressure evolution varies negligibly as a result.

The uppermost 500 m or so of the Kilauea volcano comprises some of the planet's most permeable known geological materials (with permeabilities on the order of 10^{-10} m^2 , on the basis of laboratory measurements²⁵ and simulations⁴¹). Vertical permeability, k_z , however, is as much as three orders of magnitude lower than horizontal permeability, k_x (refs. ^{24,42})—a consequence of surface-parallel layering anisotropy⁴³. Below the water table, equivalent (bulk) permeability is anticipated to be lower still. Modelling and mud-loss permeabilities⁴⁴ suggest

that k_x is in the range $1 \times 10^{-14} \text{ m}^2$ to $6 \times 10^{-14} \text{ m}^2$, with k_z estimated to be perhaps a factor of 10 or 100 lower, but not less than around 10^{-16} m^2 (ref. ⁴⁵). This is greater than laboratory measurements in this interval, partially because large-scale fractures are not encompassed by such measurements⁴³, and partially because sample recovery of friable, high-permeability materials is inherently low²⁵. Reference²⁵ reports modal porosity values of edifice-forming basalt at Kilauea of 0.15–0.3. This range is in agreement with a wealth of experimental data⁴⁶ for the typical porosity of a volcanic edifice. Thanks to these studies, we have a reasonable site-specific estimate of the permeable architecture beneath Kilauea.

Deformation experiments on edifice-forming volcanic materials⁴⁷ reveal typical values of pore compressibility (β , the inverse of the bulk modulus) on the order 10^{-10} Pa^{-1} . Accordingly, we assume a value of $K = \beta^{-1} = 10 \text{ GPa}$ throughout this study. Interstitial fluid (water) viscosity, μ , is similarly assumed to be constant ($8.9 \times 10^{-4} \text{ Pa s}^{-1}$). Permeability and porosity (the remaining parameters governing rainwater percolation through the edifice) are neither constant with depth nor independent of each other, however. Our preferred model (model α) comprises two domains in a one-dimensional half-space, the uppermost of which is highly permeable and porous ($k = 1 \times 10^{-12} \text{ m}^2$ and $\phi = 0.3$). The underlying portion, deeper than 500 m, is less permeable ($6 \times 10^{-15} \text{ m}^2$) and less porous ($\phi = 0.2$). For convenience, these data are shown in Extended Data Table 1. The results of three different additional models are shown in Fig. 3. The first is a three-section model (here called model Ω_1), which assumes that the upper 500 m of the edifice overlies intermediate- and low-permeability domains. Model Ω_2 and model Ω_3 are both homogeneous equivalent permeability models (that is, a single value of permeability is used throughout the domain), based on the parameters of model Ω_1 . Model Ω_2 reflects the arithmetic average permeability of the domains in model Ω_1 ($k_x = 8.3 \times 10^{-13} \text{ m}^2$), while model Ω_3 uses the geometric average permeability of the domains in model Ω_1 : $k_g = 5.4 \times 10^{-16} \text{ m}^2$. Values are given in Extended Data Table 1. Note that these latter models represent extreme and unlikely end-members and are shown here only for completeness.

The homogeneous models (Ω_2 and Ω_3) essentially represent upper and lower bounds of the pressure change at depth, even though the heterogeneous models may contain domains of lower equivalent permeability. Model Ω_2 exhibits very little pressure attenuation of pressure with depth and time because of the high permeability value assumed, and may be representative of an extensively fractured edifice in a self-organized critical state⁴⁸, which gives rise to hydraulic continuity from the surface to depths of several kilometres (that is, close to the pressure response of a theoretical kilometres-long vertical crack). Despite the rapid attenuation of model Ω_3 (based on the geometric average permeability), we highlight that a pressure perturbation of over 0.1 kPa is still observed at 3 km depth. The geometric average is thought to be a reasonable compromise between direction-specific averaging approaches where the precise geometry and orientation of an anisotropic medium is unknown⁴³, and so represents an approximation of a series of randomly oriented unfractured geological units. Relative to these homogeneous scenarios, the heterogeneous models (α and Ω_1) yield intermediate deviations in pressure, and comparing these models with the limited well-level data available indicate that they capture the evolution of subsurface fluid pressure in our study area (Extended Data Fig. 1). In all modelled scenarios we observe a build-up of pore pressure of tens to thousands of pascals at a depth of 3 km immediately before the onset of the 2018 rift eruption. Exploring the (k, ϕ, z) parameter space for each of the model frameworks shown in Fig. 3a reveals that this pattern holds for a wide range of feasible physical properties in a depth interval generally assumed to accommodate magma transport between Kilauea's relatively shallow magma source and the surface. In particular, this April 2018 peak is observed when the diffusivity coefficient falls within the range 0.1–1.0 for the majority of the edifice, a range in agreement with other studies that investigate fluid percolation

through volcanic media^{36,49,50} and corresponding to a wide range of physically tractable combinations of k and ϕ .

Observed well-level data

There is a paucity of available subsurface pore-pressure data for Kilauea. Reference⁵⁰ shows pore-pressure data collected for a few weeks in 2001: in this case, the time series is too short and the signal from a coincident magmatic intrusion too strong to detect the input of meteoric water. In earlier work⁵¹, pressure or head data are reported for a number of wells throughout the region around Kilauea, but are generally heavily modulated by ocean tides or located far from the rift zone. Despite these known issues, we show a time series of well-level change data digitized from ref.⁵¹, in order to highlight that a diffusion-based modelling approach is appropriate for describing groundwater evolution within the East Rift Zone (see also refs.^{50,52}). The head data are derived from aquifer tests performed in the early 1990s at the Paradise Park well (located at longitude -154.976 , latitude 19.596 ; state well number 3588-01; see Fig. 1a inset), located approximately 15 km north of the rift zone, for which there is a near-continuous record between October 1992 and September 1993. Reference⁵¹ highlights that the well is located in a portion of the aquifer that effectively dampens the tidal signal, meaning that the primary influence recorded ought to be that of the rainfall. We model pore-pressure evolution using rainfall as an input, as described in the ‘Modelling approach’ section above. In this case, we use data from the Hawai’ian Beaches rain gauge (approximately 9 km southeast from the Paradise Park well; see Fig. 1a inset), available between September 1992 and August 2005. For simplicity’s sake, we use the same input parameters as determined for Kilauea Volcano itself (based on collated experimental, modelling and drilling data), although we acknowledge that the subsurface structure at Paradise Park may differ somewhat. Because our diffusion model solely comprises physical, measurable parameters (permeability, porosity and so on), there should be no need—in theory—to tune it empirically, provided that our knowledge of the subsurface physical properties is sufficient.

Extended Data Fig. 1b shows the well-level change data⁵¹, alongside the pore-pressure change data at 1 km depth modelled here on the basis of rainfall records from the Hawai’i Beaches gauge (Extended Data Fig. 1a; see Fig. 1a inset for location). Although some discrepancies remain, clearly the modelling approach used here reproduces much of the subsurface pressure response resulting from rainfall infiltration (in particular the timing and relative magnitude of peaks, as highlighted by the grey bars). As such, we consider this model appropriate to describe subsurface pore-pressure evolution within the rift zone without empirically varying the diffusivity parameter, D .

Additional model results

Four models are described in the main text, with results shown from our preferred model, α . For reference, results from the three additional models are shown in Extended Data Fig. 2.

The parameters of model Ω_1 are based on previous experimental, numerical and in situ drilling data (see the ‘Modelling approach’ section above). Note that the maximum pore pressure occurs immediately before the 2018 fissure eruption (Extended Data Fig. 2). Based on model Ω_1 , 64% of intrusions (21 out of 33) occur when pore pressure is above the four-year average. Models Ω_2 and Ω_3 represent theoretical end-member values for hydraulic diffusivity within the East Rift Zone, and do not necessarily reflect realistic pressure-evolution scenarios. Twelve per cent of intrusions (4 of 33) occur when pore pressure is above the four-year average for Ω_2 . Fifty-five per cent of intrusions (18 of 33) occur when pore pressure is above the four-year average for Ω_3 .

Binomial probability analysis of eruption record

Fourier analysis of the satellite-derived rainfall time series (see, for example, Fig. 2) reveals that rain falls over Kilauea predominantly between 9 March and 25 August, which we define as the ‘wet’ season.

This period covers 46% of the year and accumulates >64% of annual rainfall. We analyse Kilauea’s eruption record (based on the onset date of eruptions defined in the Smithsonian Institution’s Global Volcanism Program eruption database: <https://volcano.si.edu/volcano.cfm?vn=332010>). Note that this includes the onset of the 1983 Pu’u ‘Ō’ō eruption. If eruptions were randomly (or uniformly) distributed throughout the year, we would anticipate them to occur during the ‘wet’ season with a probability of around 0.46. However, the ratio of ‘wet’-season to ‘dry’-season historical eruptions is more than 0.58. We can assess the prevalence of eruptions occurring during the ‘wet’ season by using binomial probability analysis based on the modelled time-series and Kilauea’s eruption record. The probability is calculated by:

$$\phi(x) = [n! / x! (n - x)!] \phi^x (1 - \phi)^{n-x}$$

where $\phi(x)$ is the probability of ‘successes’ out of n trials; ϕ is the probability of success of a given trial (we define a ‘success’ as an eruption that occurs within the predefined ‘rainy’ season for Kilauea); and n corresponds to the total number of observed eruptions in our dataset. This analysis allows us to determine the probability of the observed ratio of ‘wet’-season eruptions to ‘dry’-season eruptions occurring fortuitously. The mean and standard deviation of this distribution are given by $\bar{x} = n\phi$ and $\zeta = \sqrt{\bar{x}(1 - \phi)}$, respectively.

The observed value (Extended Data Fig. 3) is just within the conventional 2ζ range, which suggests that it could be a fortuitous distribution (that is, many more eruptions could have occurred during the ‘wet’ season by chance). Although this result in isolation precludes strong statistical inferences from being drawn, we highlight that the observed trend for eruptions to occur predominantly in the ‘wet’ season mirrors the prevalence of dyke intrusions occurring during periods of elevated subsurface pore pressure (approximately 60% of historical eruptions and recent dykes are associated with rainfall). We repeat the same analysis using only eruptions defined by the Global Volcanism Program as having an explosivity index value (VEI) of two or greater. Although there are too few reported larger eruptions to be statistically robust, we highlight that, again, the number of eruptions is significantly higher (to the one-sigma level) during the ‘wet’ season (four out of five VEI 2+ eruptions occurred during the ‘wet’ season; see Extended Data Fig. 3).

Data availability

Satellite-derived rainfall data (TRMM and GPM satellite data) are available from NASA’s EarthData GES DISC portal (<https://doi.org/10.5067/TRMM/TMPA/3H/7>). Rainfall gauge data are available from the NOAA’s National Centers for Environmental Information climate data portal (<https://www.ncdc.noaa.gov/cdo-web/datasets/GHCND/stations/GHCND:USC00511303/detail>). Vertical GPS data are available from the Nevada Geodetic Laboratory (<http://geodesy.unr.edu/NGLStation-Pages/stations/>; stations CRIM, AHUP, MKAI and KTPM). Additional datasets generated here are available from the corresponding author on reasonable request. Sentinel-1 ascending- and descending-track SAR acquisitions were obtained through Unavco’s Seamless SAR Archive (<https://github.com/bakerunavco/SSARA>). Vertical displacement (velocity) maps of Kilauea for the time periods 2014–2017 and 2018 are available at <https://doi.org/10.5281/zenodo.3459589>, alongside the Shuttle Radar Topography Mission (SRTM) digital elevation model used for plotting data.

Code availability

An archived version of the code required for data access, analysis and display is available at <https://doi.org/10.5281/zenodo.3635944>. Python code is in a Jupyter Notebook. Version updates, if applicable, will be made available via GitHub: https://github.com/jifarquharson/Farquharson_Amelung_2020_Kilauea-Nature and <https://github.com/>

38. Fattahi, H., Agram, P. & Simons, M. A network-based enhanced spectral diversity approach for TOPS time-series analysis. *IEEE Trans. Geosci. Remote Sens.* **55**, 777–786 (2017).
39. Yunjun, Z., Fattahi, H. & Amelung, F. Small baseline InSAR time series analysis: unwrapping error correction and noise reduction. *Computers Geosci.* **133**, 104331 (2019).
40. Blewitt, G., Hammond, W. C. & Kreemer, C. Harnessing the GPS data explosion for interdisciplinary science. *Eos* **99**, <https://doi.org/10.1029/2018EO104623> (2018).
41. Imada, J. A. *Numerical Modeling of the Groundwater in the East Rift Zone of Kilauea Volcano, Hawaii*. PhD Thesis, Univ. Hawaii at Manoa (1984).
42. Souza, W. R. & Voss, C. I. Analysis of an anisotropic coastal aquifer system using variable-density flow and solute transport simulation. *J. Hydrol.* **92**, 17–41 (1987).
43. Farquharson, J. I. & Wadsworth, F. B. Upscaling permeability in anisotropic volcanic systems. *J. Volcanol. Geotherm. Res.* **364**, 35–47 (2018).
44. Murray, J. C. *The Geothermal System at Kilauea Volcano, Hawaii*. PhD Thesis, Colorado School of Mines, Golden (1974).
45. Hsieh, P. A. & Ingebritsen, S. E. Groundwater inflow toward a preheated volcanic conduit: application to the 2018 eruption at Kilauea Volcano, Hawai'i. *J. Geophys. Res. Solid Earth* **124**, 1498–1506 (2019).
46. Farquharson, J., Heap, M. J., Varley, N. R., Baud, P. & Reuschlé, T. Permeability and porosity relationships of edifice-forming andesites: a combined field and laboratory study. *J. Volcanol. Geotherm. Res.* **297**, 52–68 (2015).
47. Heap, M. J. & Wadsworth, F. B. Closing an open system: pore pressure changes in permeable edifice rock at high strain rates. *J. Volcanol. Geotherm. Res.* **315**, 40–50 (2016).
48. Bak, P., Tang, C. & Wiesenfeld, K. Self-organized criticality: an explanation of the $1/f$ noise. *Phys. Rev. Lett.* **59**, 381 (1987).
49. Gao, S. S., Silver, P. G., Linde, A. T. & Sacks, I. S. Annual modulation of triggered seismicity following the 1992 Landers earthquake in California. *Nature* **406**, 500–504 (2000).
50. Hurwitz, S. & Johnston, M. J. Groundwater level changes in a deep well in response to a magma intrusion event on Kilauea Volcano, Hawai'i. *Geophys. Res. Lett.* **30**, 2173 (2003).

51. Gingerich, S. B. *The Hydrothermal System of the Lower East Rift Zone of Kilauea Volcano: Conceptual and Numerical Models of Energy and Solute Transport*. PhD thesis, Univ. Hawaii (1995).
52. Cervelli, P., Segall, P., Johnson, K., Lisowski, M. & Miklius, A. Sudden aseismic fault slip on the south flank of Kilauea volcano. *Nature* **415**, 1014–1018 (2002); corrigendum **418**, 108 (2002).

Acknowledgements This study would not have been possible without the Tropical Rainfall Measuring Mission (TRMM) and the Global Precipitation Measurement Mission (GPM)—joint missions between the National Aeronautics and Space Administration (NASA) and the Japanese Space Exploration Agency (JAXA)—and the European Space Agency's (ESA) Copernicus Sentinel-1 data. Copernicus Sentinel-1 data with six-day imagery are available thanks to the Group on Earth Observation's (GEO) Geohazard Supersites and Natural Laboratory Initiative (GSNL). This work was supported by funding from the NASA's Interdisciplinary Research in Earth Science (IDS) program (grant number 80NSSC17K0028 P00003). Data processing was conducted at the High Performance Computing core of the University of Miami's Center for Computational Science (CCS) using the public domain ISCE and SSARA softwares of the Jet Propulsion Laboratory (JPL) and Unavco, respectively. This study was motivated by preliminary analysis by F. Albino, and has benefited from discussions with I. Johanson, K. Anderson and D. Swanson, as well as the comments of three reviewers. Y. Zhang is thanked for his work in the development of MintPy (<https://github.com/insarlab/MintPy>). We thank NASA, the Nevada Geodetic Laboratory and the NOAA for making the data used herein freely available, as well as the developers of SSARA, ISCE and MintPy for providing free open-source software. We acknowledge Hawai'i as an indigenous space whose original people are today identified as Native Hawaiians.

Author contributions J.I.F. and F.A. conceived the study. J.I.F. performed the modelling and wrote the initial draft of the manuscript. F.A. processed the InSAR data. Both authors contributed to the discussion and interpretation of the results, and the writing of the paper.

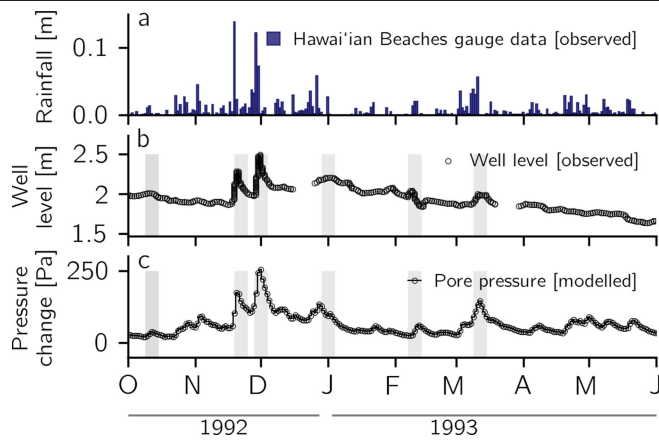
Competing interests The authors declare no competing interests.

Additional information

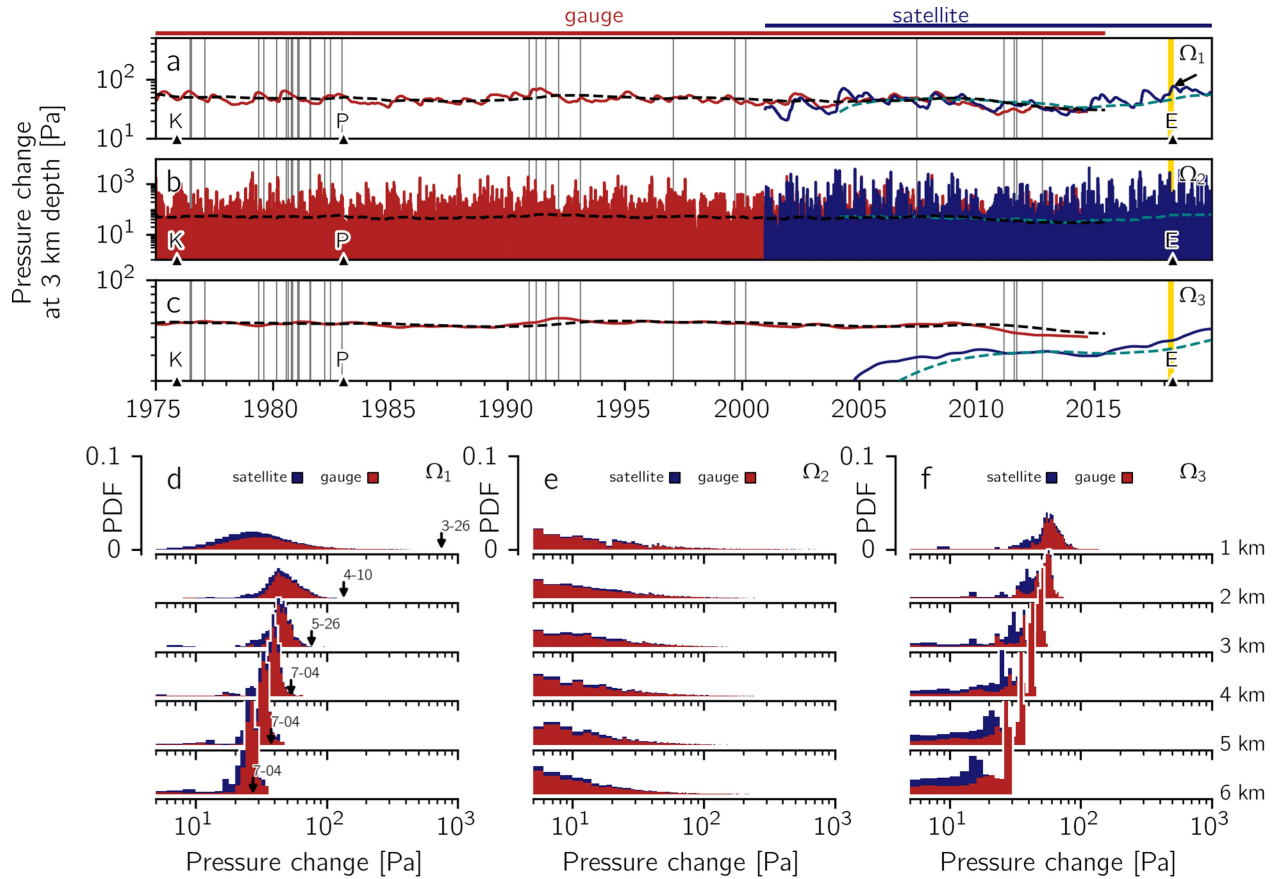
Correspondence and requests for materials should be addressed to J.I.F.

Peer review information *Nature* thanks Michael Manga and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

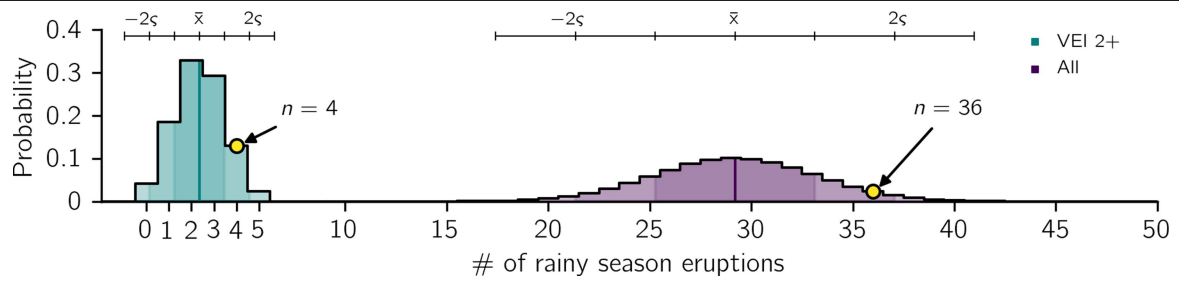


Extended Data Fig. 1 | Comparison of recorded head change and modelled pressure change. **a**, Rainfall data obtained from the Hawai‘ian Beaches rain gauge (refer to Fig. 1a inset for location). **b**, Data from the Paradise Park well (see Fig. 1a inset for the location of the well), digitized from ref.⁵¹ over the time period October 1992–June 1993. **c**, Pressure evolution at a depth of 1 km modelled using rain data from **a**. Grey bars highlight peaks evident in well-level data that are echoed in the modelled data on pore-pressure change. Note that well level serves as a proxy for pressure change, dependent on well depth and bore, inertia, storage capacity, tidal effects and atmospheric pressure: these factors are not considered in this illustrative example.



Extended Data Fig. 2 | Diffusion model results. **a**, Pore-pressure change at 3 km below the surface (1.8 km b.s.l.) modelled over the period January 1950 to April 2019 for model Ω_1 (data shown since the 1975 Kalapana earthquake). Data modelled using gauge data are shown in red; data modelled using satellite data are in blue. Four-year averages (dashed lines) are also shown. Vertical bars indicate reported intrusion events within the rift zone. K shows the Kalapana earthquake; P represents the onset of the Pu'u 'Ō'ō eruption; E highlights the 2018 fissure eruption. The arrow indicates the highest modelled pressure

change. **b**, As for **a**, but for model Ω_2 , a theoretical high-diffusivity end-member scenario. **c**, As for **a**, but for model Ω_3 , a theoretical low-diffusivity end-member scenario. **d**, PDF of modelled pressure change at depths 1–6 km below the surface from model Ω_1 . Arrows highlight the pore-pressure front diffusing from near the surface (1 km) to greater depths over time (months and dates are shown). **e**, As for **d**, but for model Ω_2 (pressure maxima not shown). **f**, As for **d**, but for model Ω_3 (pressure maxima not shown).



Extended Data Fig. 3 | Predicted binomial distribution of 'wet' season eruptions at Kilauea. The anticipated means, \bar{x} , and standard deviations, s , are shown. The observed number of historical 'wet' season eruptions (36) is highlighted, with a probability of 0.04. Data are also shown for historical eruptions of VEI2 and greater.

Extended Data Table 1 | Parameters of models shown in Fig. 3

Layer	Property	Model			
		α	Ω_1	Ω_2	Ω_3
1	$k \text{ [m}^2\text{]}$	1.0×10^{-12}	1.0×10^{-10}	8.3×10^{-12}	5.4×10^{-16}
	ϕ	0.3	0.2	0.2	0.2
	$z \text{ [km]}$	0–0.5	0–0.5	0–10	0–10
	Diffusivity $\text{[m}^2 \text{ s}^{-1}\text{]}$	3.75×10^1	5.62×10^3	4.68×10^2	3.04×10^{-2}
2	$k \text{ [m}^2\text{]}$	6.0×10^{-15}	1.0×10^{-14}	-	-
	ϕ	0.2	0.2	-	-
	$z \text{ [km]}$	0.5–10	0.5–1.2	-	-
	Diffusivity $\text{[m}^2 \text{ s}^{-1}\text{]}$	3.37×10^{-3}	5.62×10^{-1}	-	-
3	$k \text{ [m}^2\text{]}$	-	1.0×10^{-16}	-	-
	ϕ	-	0.01	-	-
	$z \text{ [km]}$	-	1.2–10	-	-
	Diffusivity $\text{[m}^2 \text{ s}^{-1}\text{]}$	-	1.12×10^{-1}	-	-

Models α and Ω_{1-3} are divided into up to up to three segments, each with a permeability, k , porosity, ϕ , and depth range, z (depth below the surface). Other parameters (K, μ) are kept constant across all models.

The projected timing of abrupt ecological disruption from climate change

<https://doi.org/10.1038/s41586-020-2189-9>

Christopher H. Trisos^{1,2,3}, Cory Merow⁴ & Alex L. Pigot⁵✉

Received: 12 January 2019

Accepted: 10 March 2020

Published online: 8 April 2020

 Check for updates

As anthropogenic climate change continues the risks to biodiversity will increase over time, with future projections indicating that a potentially catastrophic loss of global biodiversity is on the horizon^{1–3}. However, our understanding of when and how abruptly this climate-driven disruption of biodiversity will occur is limited because biodiversity forecasts typically focus on individual snapshots of the future. Here we use annual projections (from 1850 to 2100) of temperature and precipitation across the ranges of more than 30,000 marine and terrestrial species to estimate the timing of their exposure to potentially dangerous climate conditions. We project that future disruption of ecological assemblages as a result of climate change will be abrupt, because within any given ecological assemblage the exposure of most species to climate conditions beyond their realized niche limits occurs almost simultaneously. Under a high-emissions scenario (representative concentration pathway (RCP) 8.5), such abrupt exposure events begin before 2030 in tropical oceans and spread to tropical forests and higher latitudes by 2050. If global warming is kept below 2 °C, less than 2% of assemblages globally are projected to undergo abrupt exposure events of more than 20% of their constituent species; however, the risk accelerates with the magnitude of warming, threatening 15% of assemblages at 4 °C, with similar levels of risk in protected and unprotected areas. These results highlight the impending risk of sudden and severe biodiversity losses from climate change and provide a framework for predicting both when and where these events may occur.

Climate change is projected to become a leading driver of biodiversity loss¹, but it is not clear when during this century ecological assemblages might suffer such losses, and whether the process will be gradual or abrupt. Existing biodiversity forecasts typically lack the temporal perspective needed to answer these questions because they indicate the number and locations of species threatened by climate change for just a snapshot of the future, often around the end of the century^{1–3}. These snapshots do not account for the temporally dynamic nature of ecological disruption expected as a result of climate change, often focus at the level of species rather than ecological assemblages, and can seem remote to decision-makers who are concerned with managing more immediate risks⁴. Indeed, many of the most sudden and severe ecological effects of climate change can occur when conditions become unsuitable for several co-occurring species simultaneously, causing catastrophic die-offs and abrupt ‘regime shifts’ in ecological assemblages^{5,6}.

Forecasting the temporal dynamics of climate-driven disruption of ecological assemblages thus requires quantifying the differences among species in the time at which their climate niche limits may be locally exceeded. Developing advance warnings of the risk of gradual or abrupt ecological disruption is an urgent priority^{7–9}. A temporal perspective is also important for adaptation. Reducing emissions and delaying the onset of exposure to dangerous climate conditions—even

by a few decades—could buy valuable time for ecological assemblages to adapt^{10,11}, potentially reducing the magnitude of ecological disruption. However, despite the clear importance of a temporal perspective in understanding and managing the threats of climate change to biodiversity, we lack a general understanding of the time at which species in ecological assemblages will be exposed to climate conditions beyond their niche limits.

The biodiversity climate horizon

To describe the projected timing of the exposure of species to climate conditions beyond their niche, we developed an approach based on species historical climate limits and future climate projections. The range of climate conditions, over both space and time, under which a species has been recorded in the wild demarcates the boundaries of its realized niche¹². The projected time in the future at which these bounds are exceeded owing to climate change at a site can therefore be thought of as representing a climate horizon, beyond which evidence for the ability of the species to persist in the wild is lacking. Over this horizon lies, at best, a sizeable increase in uncertainty about species survival and, at worst, local extinction¹³. For a given species assemblage, the cumulative percentage of species over time that have been locally exposed to climate conditions beyond their realized niche limits forms

¹African Climate and Development Initiative, University of Cape Town, Cape Town, South Africa. ²National Socio-Environmental Synthesis Center (SESYNC), Annapolis, MD, USA. ³Centre for Statistics in Ecology, the Environment, and Conservation, University of Cape Town, Cape Town, South Africa. ⁴Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, USA.

⁵Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, London, UK. ✉e-mail: a.pigot@ucl.ac.uk

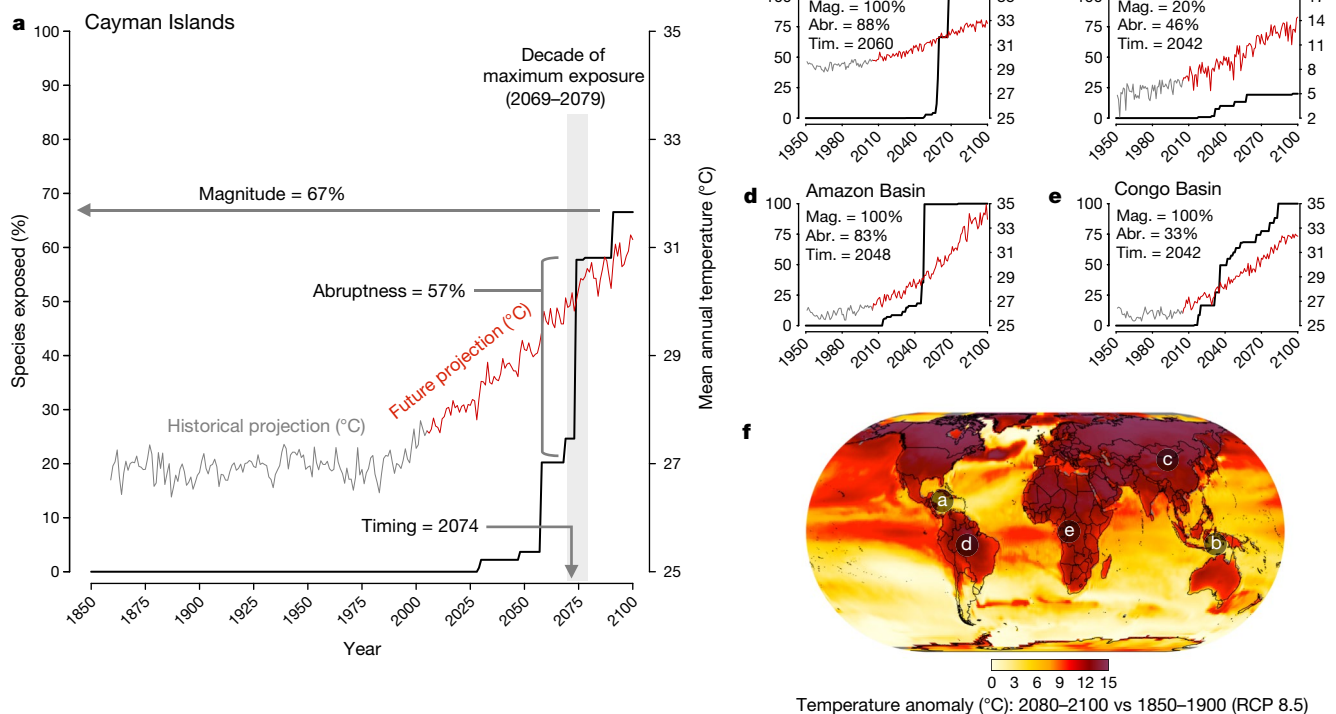


Fig. 1 | Biodiversity climate horizon profiles. a–e, Horizon profiles (solid black lines) indicate the cumulative percentage of species in an assemblage exposed to future temperatures (red lines) beyond their realized thermal niche over time. Iconic ecosystems provide examples of different profile shapes: **a**, Cayman Islands; **b**, Coral Triangle; **c**, Gobi Desert; **d**, Amazon Basin; **e**, Congo Basin. **f**, Map of temperature anomalies that shows the locations of the

ecosystems in **a–e**. Horizon profiles and temperature trends are shown for a single run of the Hadley Centre Global Environmental Model (HadGEM2) under a high greenhouse-gas-emissions scenario (RCP 8.5). The profiles differ in terms of timing, magnitude and abruptness. The grey lines show historical temperature projections at a site.

what we term the ‘horizon profile’ (Fig. 1). The shape of this horizon profile provides information on the potential for climate-driven disruption of species assemblages over time—especially the risk of early or abrupt disruption—that is not evident when focusing on individual climate snapshots.

We constructed horizon profiles for species assemblages globally, delimiting assemblages as the species occurring in 100-km grid cells based on expert-verified geographic range maps. A total of 30,652 species of birds, mammals, reptiles, amphibians, marine fish, benthic marine invertebrates, krill, cephalopods, and habitat-forming corals and seagrasses were included¹⁴ (Supplementary Table 1). We used climate projections throughout the twenty-first century from 22 climate models and 3 RCPs: strong mitigation (RCP 2.6), moderate mitigation (RCP 4.5) and a high-emissions scenario (RCP 8.5)¹⁵ (Supplementary Table 2). Given the importance of temperature as a driver of species metabolism and geographic ranges^{16–18}, we focus on mean annual temperature as the main proxy for climate. However, because species may be sensitive to other climate variables that may respond differently to greenhouse gas emissions, we also generated horizon profiles using maximum monthly temperatures and terrestrial annual precipitation (see Methods).

For each species at a site (that is, in a 100-km grid cell), we defined the local species exposure time as the year after which projected local temperatures consistently exceed—for at least 5 years—the maximum temperature experienced by the species across its geographic range during historical climate projections (1850–2005) (Supplementary Fig. 1). For species that breed annually or near-annually, 5 years represents a considerable number of breeding seasons at temperatures beyond which these species have never been recorded (a 20-year window yielded very similar results; Supplementary Figs. 2, 3). This

approach for quantifying exposure bears similarities to the concept of ‘time of emergence’ in climate science, defined as the time at which the signal of anthropogenic climate change at a location emerges from the envelope of historical climate variability^{19,20}. The key distinction is that we define exposure relative to the realized climatic niche limits of each species, rather than the historical conditions realized at a single site.

The shape of horizon profiles, and the potential ecological disruption that they imply, can vary substantially across assemblages (Fig. 1). To summarize each horizon profile, we focus on three key features: timing, the median year for an assemblage in which species exposure to unprecedented climate occurs; magnitude, the percentage of species locally exposed; and abruptness, the synchronicity in the timing of exposure among species in an assemblage, which is measured as the percentage of all species exposure times that occur in the decade of maximum exposure (Fig. 1a).

Timing, magnitude and abruptness of horizon profiles

Under RCP 8.5, 81% of terrestrial and 37% of marine assemblages are projected to have at least one species exposed to unprecedented mean annual temperatures (that is, beyond historical niche limits) before 2100. Despite the lower magnitude of warming, the magnitude of exposure is greatest in the tropics, where narrow historical climate variability²⁰ and shallow thermal gradients²¹ mean that many species occur close to their upper realized thermal limits throughout their geographic range. In total, 68% of terrestrial and 39% of tropical marine assemblages are projected to have more than 20% of their constituent species exposed to unprecedented temperatures by 2100, compared with 7% of terrestrial and 1% of marine assemblages outside the tropics (Fig. 2a). The Amazon, Indian subcontinent and Indo-Pacific regions

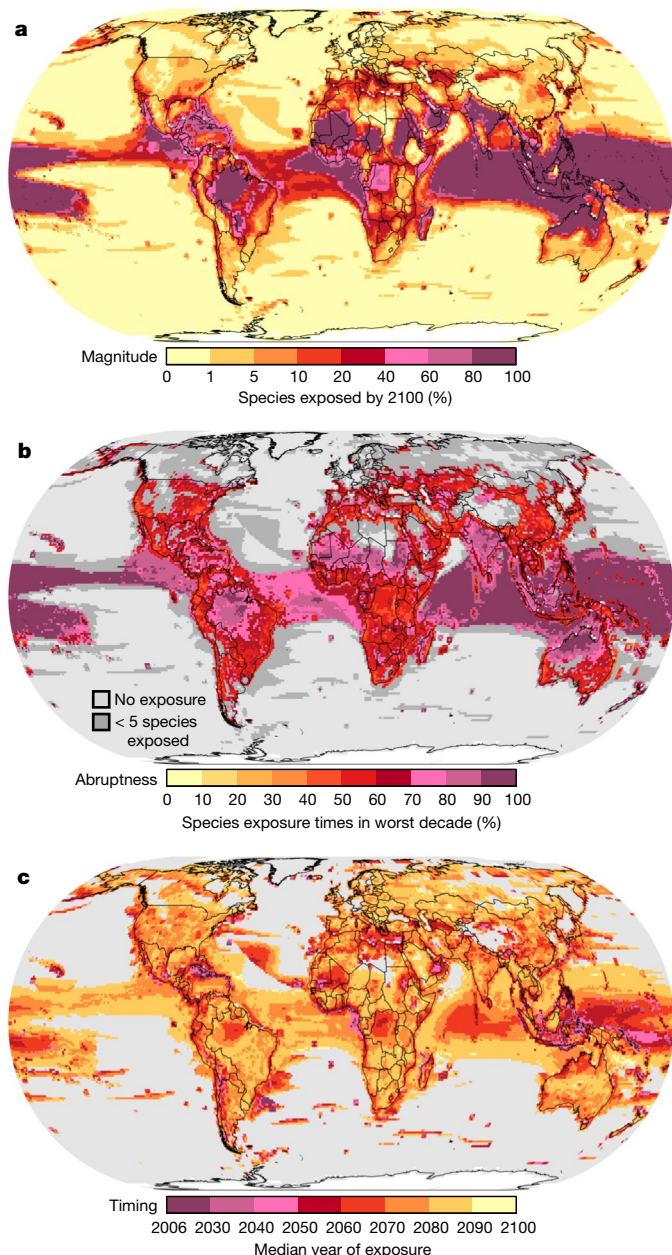


Fig. 2 | Global variation in the magnitude, abruptness and timing of horizon profiles. **a**, The magnitude of exposure is shown by the percentage of species in 100-km resolution grid cells (that is, assemblages) that are exposed to unprecedented temperature (that is, beyond the realized niche of each species) by 2100. **b**, Abruptness quantified as the percentage of species exposure times that occur within the decade of maximum exposure for each assemblage. **c**, Timing quantified as the median year of local species exposure conditional on being exposed by 2100, the end of the simulation. Maps show the median value across 22 climate models under RCP 8.5 (see Extended Data Fig. 1 for RCP 2.6 and RCP 4.5).

are most at risk, with more than 90% of species in any assemblage exposed to unprecedented temperatures by 2100 (Fig. 2a). Horizon profiles for mean annual temperature and maximum monthly temperature show strong correspondence (Extended Data Figs. 1, 2). By contrast, few species undergo prolonged exposure to unprecedented high or low annual precipitation before 2100 (Extended Data Figs. 1, 2), which is in agreement with the greater variability seen in projections of precipitation²². Thus, throughout we focus on exposure to changes in temperature.

The most notable feature of horizon profiles for local assemblages is their abruptness (Figs. 1, 2b). Under RCP 8.5, on average 71% (median) of local species exposure times for any given assemblage are projected to occur within a single decade (Fig. 3a, b), with the abruptness of exposure higher among marine assemblages (median abruptness 89%, Fig. 3a) than on land (median abruptness 61%, Fig. 3b). This pattern of highly synchronized species exposure within assemblages is robust to the choice of climate model (for RCP 8.5, median abruptness ranges from 60% to 79%; Extended Data Figs. 3, 4), emissions scenario (median abruptness 83% for RCP 2.6 and 72% for RCP 4.5), metric of abruptness (Extended Data Fig. 4), and when calculating exposure for maximum monthly temperature (median abruptness 68%) rather than mean annual temperature (Extended Data Figs. 1, 2). The same pattern of abruptness is also evident for horizon profiles constructed separately for each taxonomic group within local assemblages (Extended Data Fig. 4). Marine organisms—especially seagrasses, corals, cephalopods, marine reptiles and marine mammals—exhibit the most abrupt profiles, but it is the consistency of abruptness across groups, rather than the differences, that is most notable. Similarly, although the abruptness of exposure varies spatially—being greatest in the Amazon, Indian subcontinent, Sahel and Northern Australia, as well as tropical oceans—abrupt horizon profiles are the general rule both within the tropics (median abruptness 79%) and at higher latitudes (median abruptness 59%) (Fig. 2b).

This pervasive pattern of abrupt exposure arises primarily because co-occurring species often share similar realized thermal limits, rather than abruptness being dependent on higher rates of warming (Extended Data Fig. 5). This clustering of species-realized thermal limits can, in part, be explained by shared geographic barriers or, for tropical species, by the upper limits of temperatures available on Earth^{13,23}. However, even where these factors cannot explain the clustering of thermal limits because a high percentage of species have warmer temperatures available within 1,000 km of their range edge, assemblage exposure is still projected to occur abruptly (Extended Data Fig. 5); this suggests that other processes, such as ecological interactions²⁴ or evolutionary conservatism in fundamental niches^{25,26}, lead to similarity in realized niche limits^{16,27} and thus abruptness in the timing of exposure.

The synchronicity of species exposure within assemblages means that the timing of assemblage-level exposure events is well described by the median of species exposure times at a site (Extended Data Fig. 6). Under RCP 8.5, the global mean year of assemblage-level exposure is 2074 (± 11 years (s.d.)), but there is considerable variation in the timing of exposure across assemblages (Fig. 2c). In some locations—such as the Caribbean and the Coral Triangle—exposure is predicted to be underway already, with these hotspots of exposure expanding in spatial extent over time (Fig. 2c, Extended Data Fig. 7). By 2050, exposure spreads beyond ocean ecosystems to iconic terrestrial ecosystems, such as the Amazon, Indonesian and Congolese rainforests (Fig. 2c, Extended Data Fig. 7). Notably, the timing of these assemblage-level exposure events is not well predicted by the timing of local climate emergence (Spearman's ρ 0.29; Extended Data Fig. 5); in addition, the timing of abrupt exposure events lags behind local climate emergence by 42 years (± 12 years; mean \pm s.d.), indicating the potential time-lag between climate change and ensuing biotic responses.

The abrupt exposure of species within ecological assemblages has not been detected in earlier projections of climate-driven range loss and global species extinctions, which have implied a more gradual increase in risk to biodiversity^{2,3}. We find that the appearance of a gradual increase in risk can result from summarizing across local assemblages that differ in their projected timing of abrupt exposure (Fig. 3c, d, Extended Data Fig. 8). Although these global summaries mask the abrupt nature of exposure within local assemblages, they can highlight the importance of increased mitigation efforts in reducing and delaying the onset of unprecedented climate conditions. Compared to RCP 8.5, achieving RCP 2.6 delays exposure for the most at-risk species

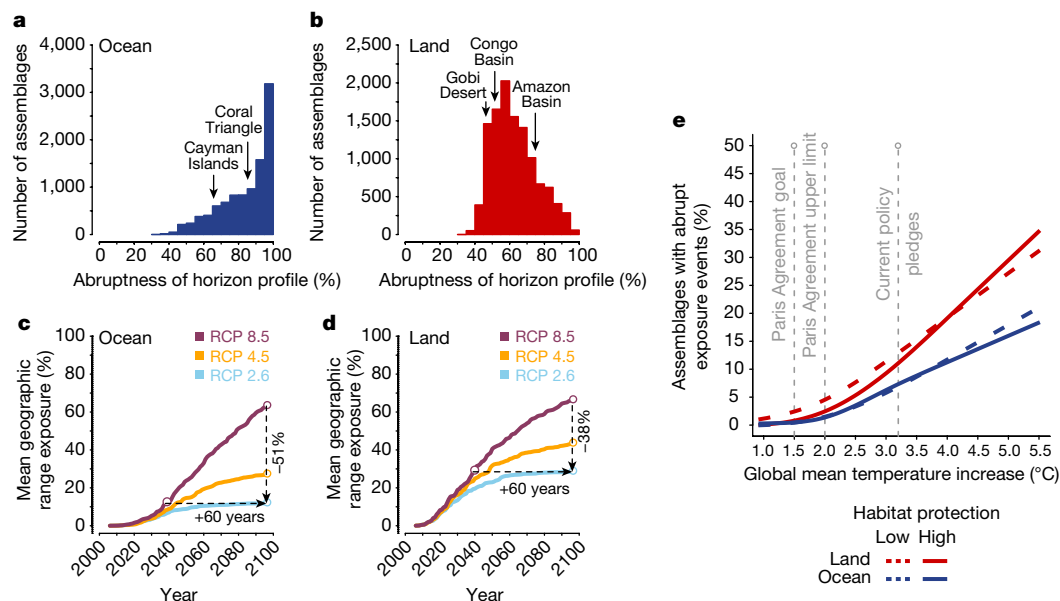


Fig. 3 | Abruptness of horizon profiles locally compared with globally, and the accelerating risk with global warming. **a, b,** The distribution in the projected abruptness of species exposure to unprecedented temperatures within marine (**a**) and terrestrial (**b**) assemblages. Selected assemblages from Fig. 1 are highlighted. Abruptness is quantified as the percentage of species exposure times that occur within the decade of maximum exposure, with results showing the median across climate models under RCP 8.5. **c, d,** Global horizon profiles for oceans (**c**) and land (**d**) show more gradual accumulation of species exposure to unprecedented temperatures. Dashed lines show how lowering emissions from RCP 8.5 to RCP 2.6 both reduces the median

by approximately 6 decades in the oceans (mean 58 years, range 46–65 years; Fig. 3c) and on land (mean 58 years, range 49–67 years; Fig. 3d), buying valuable time for species and ecosystems—and human societies that depend on them—to adapt to a warming climate.

The risk of abrupt exposure events

The abruptness of horizon profiles is positively correlated with the magnitude of exposure (Spearman's ρ 0.58; Extended Data Fig. 6), which indicates that exposure events involving larger fractions of species are projected to occur more abruptly. This near-simultaneous exposure among multiple species could have sudden and devastating effects on local biodiversity and ecosystem services. Catastrophic, multi-species coral die-offs caused by a record-breaking marine heatwave in 2016 are one recent example⁶. Although the 'safe limits' of species loss—at which ecosystem function can be maintained—remain uncertain, meta-analyses suggest a 20% decline in species diversity as one possible threshold^{28,29}. We therefore defined assemblages at risk of abrupt ecological disruption as those in which at least 20% of species are projected to undergo exposure to unprecedented temperatures within the same decade. Restricting global warming to less than 2 °C above pre-industrial levels limits such abrupt assemblage exposure events to less than 2% of assemblages (Fig. 3e). However, beyond 2 °C warming, the area projected to undergo abrupt assemblage exposure expands rapidly, encompassing 15% of assemblages globally at 4 °C warming. Furthermore, the increase in abrupt exposure does not differ markedly for assemblages that are afforded high habitat protection (at least 20% protected area coverage of a grid cell), indicating that current protected areas are equally at risk from abrupt exposure (Fig. 3e).

The risk of abrupt exposure events differs across assemblages globally, with variability across individual climate projections increasing the total area at risk compared with median projections. For instance,

magnitude of exposure across climate models and substantially delays the timing of exposure, buying about 60 years for species and conservation plans to adapt to a warming climate (see Extended Data Fig. 8 for individual climate models). **e,** The percentage of species assemblages projected to experience high-magnitude and abrupt assemblage exposure (more than 20% of species exposed in a single decade) as a function of global mean temperature increase relative to pre-industrial levels (1850–1900). Curves are fitted from model runs ($n = 66$) across RCP 2.6, RCP 4.5 and RCP 8.5. 'Current policy pledges' refers to a scenario in which countries implement their current nationally determined contributions to 2030 and make no further emissions reductions.

even under RCP 2.6 (1.75 °C mean warming), 9% of assemblages are at some risk of abrupt exposure (Fig. 4a), and this increases to 35% of assemblages under RCP 8.5 (4.4 °C mean warming; Fig. 4b). The risk of abrupt assemblage exposure events is positively correlated with species richness (RCP 8.5, Spearman's ρ 0.29 (land) and 0.56 (ocean)), highlighting the increased risk of sudden ecological disruption in the world's most biodiverse ecosystems. Moreover, the risk of disruption of ecological function may be underestimated in this analysis because even if particular functional groups (for example, habitat-forming corals) suffer high levels of exposure, this may not be evident at the scale of entire assemblages if other groups are relatively less affected. When abrupt assemblage exposure events are instead defined at the level of major taxonomic groups, the area at risk expands further, encompassing 49% of species assemblages under RCP 8.5 (Fig. 4c, Extended Data Fig. 9). Our approach estimates how much of the original biodiversity of an assemblage is exposed to potentially dangerous climate conditions over time²⁸. We do not consider the potential for immigration of species from elsewhere to offset local biodiversity losses; however, abrupt assemblage-wide exposure is likely to cause substantial ecological disruption regardless of the rate at which new species arrive. Furthermore, in tropical lowlands and oceans—where projected exposure is greatest and species adapted to warmer environments are lacking—net declines in local biodiversity are expected²¹.

Crossing the biodiversity climate horizon

Although the horizon profile describes the accumulating number of co-occurring species that are exposed to conditions beyond their realized niche limits, this need not equate with a profile of local extinction. Species may have wider fundamental niche limits than realized niche limits^{13,30}, may avoid exposure in microclimatic refugia (however, see Extended Data Fig. 10) or through behavioural thermoregulation^{17,31},

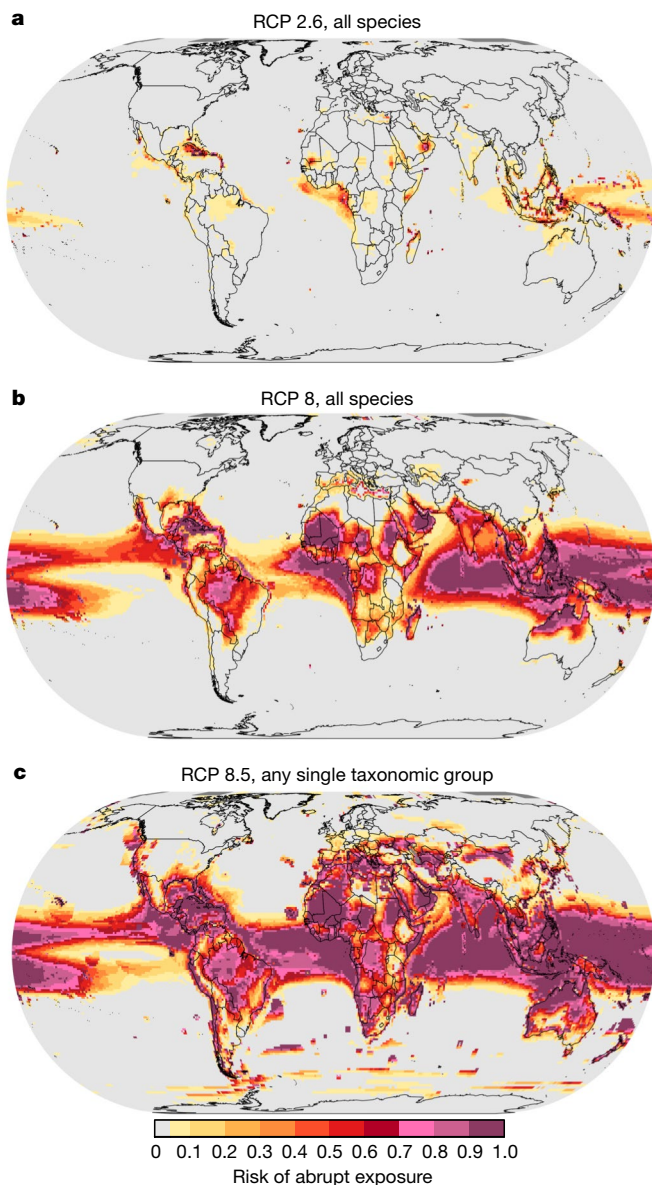


Fig. 4 | The risk of high-magnitude, abrupt assemblage exposure events. **a–c**, Risk is shown for all species under RCP 2.6 (**a**), all species under RCP 8.5 (**b**) and single taxonomic groups under RCP 8.5 (**c**). Risk is calculated as the proportion of 22 climate models in which an abrupt exposure event is projected to occur before 2100. Assemblages that avoid abrupt exposure events across all 22 models are in grey. In **a, b**, abrupt exposure events are defined as when more than 20% of all species in an assemblage are exposed in a single decade. In **c**, abrupt exposure events are defined when any single group of organisms (for example, amphibians or corals) within an assemblage experiences the exposure of more than 20% of its constituent species in a single decade, highlighting the widespread risk of abrupt ecological disruption.

or may evolve to tolerate new conditions¹⁰. In these cases, the timing of abrupt assemblage exposure events could be considered an ‘ignorance horizon’—the time beyond which local extinctions are not inevitable but evidence for the ability of species to persist in the wild is largely absent¹³. Thus, at the very least, our results show that within 30 years, continued high emissions will drive a sudden shift across many ecological assemblages to climate conditions under which we have almost no knowledge of the ability of their constituent species to survive. We caution that the timing and magnitude of this exposure may occur earlier and be larger than we anticipate, because our analysis does not consider changes in extreme events⁹, effects of warming on local

habitat (for example, melting sea ice), covariation between climate variables³², or that populations may be locally adapted³³.

Furthermore, to the extent that species-realized historical thermal limits do reflect fundamental limits to persistence, then the occurrence of abrupt exposure events marks the crossing of an ‘ecological horizon’ beyond which catastrophic and coordinated species losses are expected. These abrupt events—projected to spread from ocean (for example, coral reef) to land (for example, rainforest) ecosystems by 2050 under high emissions—risk sudden disruption to ecosystems and their capacity to maintain current levels of biodiversity and functioning. Evidence from laboratory-based and field-based studies indicates this is a credible risk, particularly for tropical terrestrial ectotherms and for marine organisms for which projected abruptness is most pronounced and for which realized geographic range boundaries most closely match thermal tolerance limits^{16,18,30,34}. Indeed, warming over recent decades has already been associated with marked population declines and local extinctions^{6,35,36}—even among endotherms, which are widely assumed to be less sensitive to warming but may be particularly vulnerable to climate-driven disruption of trophic interactions^{37,38}. For those ecosystems for which exposure is projected within the next few decades, the capacity for species to adapt would appear limited. A priority for future research is to refine estimates of the timing and consequences of exposure, including for regions in which factors other than temperature may more strongly constrain species ranges, and for which the emergence of novel climates has closest analogues deep in Earth’s history³⁹.

Considering the temporal dynamics of the exposure of biodiversity to climate change provides an early warning of the potential for abrupt ecological disruption. Averting—or at least delaying—the crossing of this ecological horizon is possible for most assemblages, and requires massive and rapid reductions in greenhouse gas emissions. Our results also highlight the urgency of targeted management responses, including establishing monitoring sites in exposed regions, establishing new protected areas in refugia, and investigating the potential of assisted migration and adaptation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2189-9>.

- Urban, M. C. Accelerating extinction risk from climate change. *Science* **348**, 571–573 (2015).
- Warren, R., Price, J., Graham, E., Forstenhaeusler, N. & VanDerWal, J. The projected effect on insects, vertebrates, and plants of limiting global warming to 1.5 °C rather than 2 °C. *Science* **360**, 791–795 (2018).
- Newbold, T. Future effects of climate and land-use change on terrestrial vertebrate community diversity under different scenarios. *Proc. R. Soc. B* **285**, 20180792 (2018).
- Weber, C. et al. Mitigation scenarios must cater to new users. *Nat. Clim. Change* **8**, 845–848 (2018).
- Wernberg, T. et al. Climate-driven regime shift of a temperate marine ecosystem. *Science* **353**, 169–172 (2016).
- Hughes, T. P. et al. Global warming transforms coral reef assemblages. *Nature* **556**, 492–496 (2018).
- Barnosky, A. D. et al. Approaching a state shift in Earth’s biosphere. *Nature* **486**, 52–58 (2012).
- Scheffer, M. et al. Early-warning signals for critical transitions. *Nature* **461**, 53–59 (2009).
- Harris, R. M. B. et al. Biological responses to the press and pulse of climate trends and extreme events. *Nat. Clim. Change* **8**, 579–587 (2018).
- Bay, R. A., Rose, N. H., Logan, C. A. & Palumbi, S. R. Genomic models predict successful coral adaptation if future ocean warming rates are reduced. *Sci. Adv.* **3**, e1701413 (2017).
- Chevin, L.-M., Lande, R. & Mace, G. M. Adaptation, plasticity, and extinction in a changing environment: towards a predictive theory. *PLoS Biol.* **8**, e1000357 (2010).
- Colwell, R. K. & Rangel, T. F. Hutchinson’s duality: the once and future niche. *Proc. Natl Acad. Sci. USA* **106**, 19651–19658 (2009).
- Feeley, K. J. & Silman, M. R. Biotic attrition from tropical forests correcting for truncated temperature niches. *Glob. Change Biol.* **16**, 1830–1836 (2010).
- The IUCN Red List of Threatened Species <https://www.iucnredlist.org/> (IUCN, 2017).

15. van Vuuren, D. P. et al. The representative concentration pathways: an overview. *Clim. Change* **109**, 5–31 (2011).
16. Stuart-Smith, R. D., Edgar, G. J. & Bates, A. E. Thermal limits to the geographic distributions of shallow-water marine species. *Nat. Ecol. Evol.* **1**, 1846–1852 (2017).
17. Sunday, J. M. et al. Thermal-safety margins and the necessity of thermoregulatory behavior across latitude and elevation. *Proc. Natl Acad. Sci. USA* **111**, 5610–5615 (2014).
18. Dillon, M. E., Wang, G. & Huey, R. B. Global metabolic impacts of recent climate warming. *Nature* **467**, 704–706 (2010).
19. Hawkins, E. & Sutton, R. Time of emergence of climate signals. *Geophys. Res. Lett.* **39**, L01702 (2012).
20. Mora, C. et al. The projected timing of climate departure from recent variability. *Nature* **502**, 183–187 (2013).
21. Colwell, R. K., Brehm, G., Cardelús, C. L., Gilman, A. C. & Longino, J. T. Global warming, elevational range shifts, and lowland biotic attrition in the wet tropics. *Science* **322**, 258–261 (2008).
22. IPCC. *Climate Change 2013: The Physical Science Basis* (eds Stocker, T. F. et al.) (Cambridge Univ. Press, 2013).
23. Williams, J. W., Jackson, S. T. & Kutzbach, J. E. Projected distributions of novel and disappearing climates by 2100 AD. *Proc. Natl Acad. Sci. USA* **104**, 5738–5742 (2007).
24. Liautaud, K., van Nes, E. H., Barbier, M., Scheffer, M. & Loreau, M. Superorganisms or loose collections of species? A unifying theory of community patterns along environmental gradients. *Ecol. Lett.* **22**, 1243–1252 (2019).
25. Araújo, M. B. et al. Heat freezes niche evolution. *Ecol. Lett.* **16**, 1206–1219 (2013).
26. Crisp, M. D. et al. Phylogenetic biome conservatism on a global scale. *Nature* **458**, 754–756 (2009).
27. White, A. E., Dey, K. K., Mohan, D., Stephens, M. & Price, T. D. Regional influences on community structure across the tropical–temperate divide. *Nat. Commun.* **10**, 2646 (2019).
28. Newbold, T. et al. Has land use pushed terrestrial biodiversity beyond the planetary boundary? A global assessment. *Science* **353**, 288–291 (2016).
29. Hooper, D. U. et al. A global synthesis reveals biodiversity loss as a major driver of ecosystem change. *Nature* **486**, 105–108 (2012).
30. Sunday, J. M., Bates, A. E. & Dulvy, N. K. Thermal tolerance and the global redistribution of animals. *Nat. Clim. Change* **2**, 686–690 (2012).
31. Pinsky, M. L., Eikeset, A. M., McCauley, D. J., Payne, J. L. & Sunday, J. M. Greater vulnerability to warming of marine versus terrestrial ectotherms. *Nature* **569**, 108–111 (2019).
32. Mahony, C. R. & Cannon, A. J. Wetter summers can intensify departures from natural variability in a warming climate. *Nat. Commun.* **9**, 783 (2018).
33. Valladares, F. et al. The effects of phenotypic plasticity and local adaptation on forecasts of species range shifts under climate change. *Ecol. Lett.* **17**, 1351–1364 (2014).
34. Deutsch, C. A. et al. Impacts of climate warming on terrestrial ectotherms across latitude. *Proc. Natl Acad. Sci. USA* **105**, 6668–6672 (2008).
35. Sinervo, B. et al. Erosion of lizard diversity by climate change and altered thermal niches. *Science* **328**, 894–899 (2010).
36. Soroye, P., Newbold, T. & Kerr, J. Climate change contributes to widespread declines among bumble bees across continents. *Science* **367**, 685–688 (2020).
37. Lister, B. C. & Garcia, A. Climate-driven declines in arthropod abundance restructure a rainforest food web. *Proc. Natl Acad. Sci. USA* **115**, E10397–E10406 (2018).
38. Spooner, F. E. B., Pearson, R. G. & Freeman, R. Rapid warming is associated with population decline among terrestrial birds and mammals globally. *Glob. Change Biol.* **24**, 4521–4531 (2018).
39. Burke, K. D. et al. Pliocene and Eocene provide best analogs for near-future climates. *Proc. Natl Acad. Sci. USA* **115**, 13288–13293 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Biodiversity data

We used expert-verified range maps for 30,652 species from the International Union for Conservation of Nature (IUCN)¹⁴ and BirdLife International⁴⁰, including birds, mammals, reptiles, amphibians, marine fish, benthic marine invertebrates, and habitat-forming corals and seagrasses (Supplementary Table 1). To further increase coverage of open-ocean assemblages, our sample includes additional data on krill⁴¹ and cephalopods⁴², reflecting the availability of expert range maps for oceanic species⁴³. We used only breeding ranges for terrestrial species. We excluded marine species that are restricted to depths greater than 200 m (the lower limit of the epipelagic zone), as these species are less likely to respond to changes in sea surface temperature. Range maps were converted to 100-km-resolution equal-area grid cells—the finest resolution justifiable for these data globally without incurring false presences^{44,45}. Expert range maps provide comprehensive information on the global geographic distributions of species⁴⁶, but our results should be interpreted in the context of known data limitations. For some groups, species coverage is incomplete and is biased towards commercial species (for example, cephalopods), whereas others have been comprehensively assessed for only a subset of clades (for example, fish) and the species included in our study thus represent a non-random subset of global biodiversity (Supplementary Table 1). For instance, insects and plants may on average be more at risk of geographic range loss due to climate change than are terrestrial vertebrates², but we did not assess exposure for these groups because range maps (expert or otherwise) are not available globally. As such, both very short-lived and long-lived terrestrial taxa may be underrepresented in our sample. Furthermore, although many IUCN range maps consider occurrence data from historical records, others may underestimate climate niche limits where longer-term historical records are unavailable and recent geographic range contractions have occurred in part due to reasons other than climate change⁴⁷.

Data on marine and terrestrial protected areas were downloaded from the World Database on Protected Areas (<http://protectedplanet.net/>; accessed 21 March 2018). The maps, originally in polygon format, were resampled to a 1-km resolution before further analysis. We considered 100-km-resolution grid cells highly protected if at least 20% of the grid cell was inside protected areas.

Climate model data

We used temperature and precipitation projections from 22 General Circulation and Earth System Models developed for the Coupled Model Intercomparison Project 5 (CMIP5) (Supplementary Table 2). For each model, we downloaded a single projection for mean monthly precipitation (mm), near-surface temperature (K) and sea surface temperature (K) for the historical run (1850–2005), as well as RCP 2.6, RCP 4.5 and RCP 8.5 scenarios for the years 2006–2100 or 2006–2300, when available. Model output was downloaded from <https://esgf-node.llnl.gov/projects/esgf-llnl/> (accessed 5 June 2017). In our main analysis, we focus on the dynamics of exposure according to mean annual temperature (MAT), calculated by averaging monthly values. However, we also repeated our analysis using the temperature of the hottest month, hereafter denoted maximum monthly temperature (MMT), and, for terrestrial assemblages, total annual precipitation (mm), calculated by summing precipitation values across months (see Supplementary Information). Note that the identity of the hottest month can vary both across sites and between years within a site. Given that CMIP5 models use different spatial grids, and to match the resolution of species geographic range data, climate model data were regridded to a 100-km resolution grid using an area-weighted mean interpolation. Climate data interpolation was performed in CDO⁴⁸ and R⁴⁹.

We calculated species exposure times for each assemblage using individual climate simulations, as opposed to ensembles or multi-model

averages, because individual simulation runs include variance in climatic time series due to internal climate variability such as the timing of El Niño/Southern Oscillation events^{22,50}. This internal variability is a key component of the uncertainty in the timing of exposure, and is smoothed out if using multi-model averages as input into the analysis. By calculating species exposure events using individual model simulation runs and then summarizing across models, we capture the uncertainty in the timing of exposure due to both internal climate variability and climate model uncertainty (that is, uncertainty about climate physics across models), in line with ‘time of emergence’ analyses from climate science¹⁹. Throughout, we report multi-model medians in each of our summary metrics.

Defining species-realized niche limits

Species experience variability in climatic conditions across both space and time, but this temporal variability is ignored when using time-averaged climate conditions (for example, Worldclim data⁵¹) to estimate species-realized niches. To address this, we estimated species-realized niche limits using the climate projections from the historical run of each climate model (1850–2005), which includes the influence on climate of observed changes in radiative forcing due to natural factors such as volcanic eruptions, as well as anthropogenic emissions and land-use changes⁵². Thus, in the case of MAT, we calculated the maximum MAT experienced across the species geographic range over both space and time ($T_{\max_{\text{MAT}}}$, see Supplementary Information). To prevent estimates of $T_{\max_{\text{MAT}}}$ being inflated by either extreme outliers in the temperature time series or from the overestimation of species ranges⁴⁴, we excluded outlier temperature values within each grid cell, defined as those more than three standard deviations from the mean. After we had selected the maximum temperature for each cell, we excluded outlier temperature values across each species range, defined as those more than three standard deviations above the mean range value. The $T_{\max_{\text{MAT}}}$ value for each species was then set as the maximum of the remaining values (Supplementary Fig. 1). We used an identical procedure to calculate T_{\max} using MMT ($T_{\max_{\text{MMT}}}$). For precipitation, species may be exposed to either drying or wetting conditions and so we calculated both the minimum (P_{\min}) and maximum (P_{\max}) precipitation values experienced by each species across its geographic range (see Supplementary Information).

Estimating species exposure times

Within each terrestrial ($n = 18,560$) and marine ($n = 37,333$) assemblage (that is, 100-km grid cell containing any terrestrial or marine species, respectively), we defined the time of local species exposure to unprecedented temperature (that is, the climate horizon) to be the year after which the MAT (or MMT) of the cell is projected to exceed the $T_{\max_{\text{MAT}}}$ (or $T_{\max_{\text{MMT}}}$) value of the species for at least five consecutive years. We note that using a higher number of consecutive years ($n = 20$ years) had little effect on the magnitude, timing or abruptness of exposure (Supplementary Figs. 2, 3).

For precipitation, we calculated the time of local species exposure as the year after which the precipitation of the cell is projected to be either greater or less than the P_{\max} and P_{\min} values, respectively, of the species for at least five consecutive years. Annual precipitation values are bounded at zero, and this could potentially lead to exposure being underestimated for locations projected to have historically received zero precipitation. To address this, we additionally defined exposure to occur when annual precipitation decreased to less than 15 mm for at least five consecutive years. Owing to the generally weaker trends and high variability in historical and future projected precipitation, we found that few species were exposed to unprecedented precipitation regardless of how exposure was defined (Extended Data Fig. 1). To show the importance of increasing temperatures as the primary driver of exposure, we compared patterns of exposure from MAT alone to those from MAT and precipitation combined, recording the earliest

local exposure time of either MAT or precipitation for a species in an assemblage when it was exposed to both variables (Extended Data Fig. 2, see Supplementary Information).

We note that by using range-wide estimates of species niche limits, we may underestimate both the magnitude and the immediacy of exposure if populations are locally adapted³³. Unfortunately, information on the scale and strength of local adaptation is not generally available across species. Equally, our analysis does not attempt to model adaptive evolution, which may enable species to shift or expand their climatic niche limits over time. Nevertheless, our estimates of the timing of local exposure to unprecedented conditions may be relevant for understanding the potential for evolution to rescue populations from changing climates^{10,11}.

Horizon profiles

When species exposure times had been calculated for an assemblage, we constructed a horizon profile indicating the cumulative percentage of species that are locally exposed to conditions beyond their realized niche limits. We used the following metrics to summarize the temporal dynamics of biodiversity exposure. First, we calculated the magnitude of exposure as the percentage of species in the assemblage exposed over the course of the twenty-first century. Second, the abruptness of exposure for an assemblage was calculated as the percentage of all exposure times that occur in the decade of maximum exposure. We identified the decade of maximum exposure using a moving window of ten years. We also calculated an alternative metric of abruptness based on the Shannon entropy index⁵³, which quantifies the evenness in the distribution of exposure times across all decades of the horizon profile (Extended Data Fig. 4). In contrast to our original abruptness metric, lower values of the Shannon entropy index indicate a more abrupt profile. We therefore rescaled the Shannon entropy index by the maximum possible entropy value per assemblage, subtracted these values from 1 and then multiplied by 100 to give an index that ranged between 0 and 100, where a value of 100 indicates that all exposure times occur in a single decade and a value of 0 corresponds to an equitable distribution of exposure times across decades. Abruptness was calculated only for assemblages in which five or more species were exposed, to avoid idiosyncrasies due to small sample sizes. Third, the timing of exposure for each assemblage was calculated as the median of the times of local species exposure events. Species not exposed before the end of the twenty-first century were excluded from this calculation. We repeated our analysis using alternative metrics of timing, including the mean year of exposure and the mid-point of the decade of maximum exposure, obtaining very similar results (Extended Data Fig. 6). For each of these exposure metrics we report the median value across the 22 climate models for a given climate scenario, and quantify uncertainty as the standard deviation (Extended Data Fig. 3). The greatest uncertainty in projected effects involves the magnitude of exposure along the boundaries of the tropics. This arises because of variation among models in the magnitude of warming, which alters the spatial extent of regions exposed to unprecedented temperatures. By contrast, variation among models in the timing and abruptness of exposure is relatively small and does not exhibit any clear spatial structure.

We compared the median timing of species exposure within assemblages to the timing of local climate emergence, defined as the year after which future local temperatures are projected to exceed the maximum historical (1850–2005) conditions at a site^{19,20}. Timing of emergence was calculated using an identical procedure to that used to calculate the timing of exposure, excluding outlying values from the time series when quantifying the maximum historical temperature at a site and considering emergence only when temperatures exceed the historical maximum for at least five consecutive years. The time of local climate emergence at a site is therefore identical to the time of local exposure for a species occupying a single grid cell. In the absence of perfect adaptation to local climates, a time-lag is therefore expected between

local climate emergence and the median timing of exposure, because species typically persist under a broader range of conditions than is present in any single site.

Spatial scale

We modelled species-realized niche limits using climate projections at 100-km grain size, matching the resolution of expert geographic range maps^{44,45}. However, individual grid cells at this resolution may contain (potentially substantial) spatial climatic heterogeneity, thus potentially underestimating variability in species niche limits and potentially overestimating the abruptness of assemblage exposure dynamics. To investigate this possibility, we tested whether the abruptness of horizon profiles across terrestrial assemblages is related to the range in the MAT within each grid cell, using spatially interpolated temperature data for the period 1970–2000 available at 1-km resolution⁵¹. We found that abruptness is negatively correlated with the spatial heterogeneity in temperature within a cell (Spearman's $\rho = -0.29$), so that assemblages with higher spatial heterogeneity in temperatures (for example, tropical mountains), exhibit more gradual exposure profiles than those with low heterogeneity in temperatures (for example, tropical lowlands) (Extended Data Fig. 10). This result has two important implications. First, it suggests that—despite the relatively coarse grain size—our analysis still identifies those assemblages in which variation in realized niche limits among species is expected to be greatest (that is, grid cells containing substantial spatial climatic heterogeneity) as having the most gradual exposure profiles. Second, it suggests that, although incorporating finer-scale climate data may further reduce the lowest abruptness values estimated across assemblages (that is, making relatively gradual horizon profiles more gradual), it is unlikely to alter the key conclusion that assemblage exposure to climate warming occurs abruptly, because the most abrupt horizon profiles occur in assemblages in which there is little fine-scale climatic heterogeneity (Extended Data Fig. 10). These results support the robustness of our overall conclusions regarding the dynamics of exposure, but it is clear that increasing the spatial resolution at which species niche limits and assemblages are defined would enable a more precise quantification of the timing of species exposure to changing climates, and should be a priority for future research.

Horizon profiles can be calculated either for a single assemblage or for a set of assemblages combined, such as a biome or the entire globe. In addition to examining the dynamics within assemblages, we generated global horizon profiles, describing the total cumulative exposure of all populations (that is, species by site combinations) across marine and terrestrial assemblages (Fig. 3c, d). To avoid exposure dynamics being driven by the small number of species with the largest geographic ranges, we weighted each species by the inverse of its geographic range size. This range-size-weighted exposure profile ensures that each species contributes equally to exposure dynamics, and is mathematically equivalent to calculating the mean percentage geographic range exposure across species. Unweighted global horizon profiles show qualitatively similar patterns (Extended Data Fig. 8).

Risk of abrupt exposure events

We identified those assemblages projected to undergo abrupt and high-magnitude exposure events, defined as at least 20% of resident species exposed within a single decade before the end of the twenty-first century. Across the set of 66 climate model runs from the 3 RCP scenarios, we fit a generalized additive model to estimate the percentage of assemblages projected to undergo abrupt exposure events as a function of mean global warming at the end of the century (2080–2100) relative to pre-industrial conditions (1850–1900). We fit separate models for sites with either low or high (that is, greater than 20% in protected areas) levels of habitat protection. We forced the regression through the origin, thus assuming no abrupt exposure events would occur if temperatures remained stable at pre-industrial conditions. Because

Article

the identity of assemblages projected to undergo abrupt exposure events may vary across model runs, the actual area at risk of abrupt exposure may be substantially greater than expected under any single climate simulation. For each assemblage, we therefore calculated the probability of an abrupt exposure event across the 22 climate models within each emissions scenario. We did this for assemblages consisting of all species, as well as for each group of organisms separately.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All datasets used here are publicly available. Expert verified range maps are available from <https://www.iucnredlist.org/resources/spatial-data-download> and <http://datazone.birdlife.org/species/requestdis>. Climate change projections for RCP 8.5, RCP 4.5 and RCP 2.6 for CMIP5 are available from <https://esgf-node.llnl.gov/search/cmip5/>. Maps of projected risk to biodiversity from climate change are available to view at <https://climatehorizons.users.earthengine.app/view/biodiversity-risk>.

Code availability

Computer code used in the analysis is available on request from the corresponding author. Code and data that were used to make Figs. 2–4 is available at Figshare (<https://doi.org/10.6084/m9.figshare.11814633>).

40. Bird Species Distribution Maps of the World v.2.0 (Birdlife International, 2012).
41. Brinton, E., Ohman, M. D., Townsend, A. W., Knight, M. D. & Bridgeman, A. L. *Euphausiids of the World Ocean* (Springer, 2000).
42. Jereb, P. & Roper, C. F. E. (eds) *Cephalopods of the World: An Annotated and Illustrated Catalogue of Cephalopod Species Known to Date* Vol. 1 (FAO, 2005).
43. Tittensor, D. P. et al. Global patterns and predictors of marine biodiversity across taxa. *Nature* **466**, 1098–1101 (2010).

44. Hurlbert, A. H. & Jetz, W. Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. *Proc. Natl Acad. Sci. USA* **104**, 13384–13389 (2007).
45. Jetz, W., Sekercioglu, C. H. & Watson, J. E. M. Ecological correlates and conservation implications of overestimating species geographic ranges. *Conserv. Biol.* **22**, 110–119 (2008).
46. Meyer, C., Kreft, H., Guralnick, R. & Jetz, W. Global priorities for an effective information basis of biodiversity distributions. *Nat. Commun.* **6**, 8221 (2015).
47. Faurby, S. & Araújo, M. B. Anthropogenic range contractions bias species climate change forecasts. *Nat. Clim. Change* **8**, 252–256 (2018).
48. Schulzweida, U. CDO User Guide v1.9.6 <https://doi.org/10.5281/zenodo.2558193> (2019).
49. R Core Team. R: a language and environment for statistical computing. <http://www.R-project.org/> (R Foundation for Statistical Computing, 2019).
50. Kay, J. E. et al. The Community Earth System Model (CESM) Large Ensemble Project: a community resource for studying climate change in the presence of internal climate variability. *Bull. Am. Meteorol. Soc.* **96**, 1333–1349 (2015).
51. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978 (2005).
52. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
53. Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 623–656 (1948).

Acknowledgements We thank G. Mace and O. Petchey for comments on pre-submission drafts of the manuscript. This study has been supported by the following institutions and grants: the Royal Society, UK, to A.L.P.; the National Socio-Environmental Synthesis Center under funding received from the National Science Foundation DBI-1639145 and the FLAIR Fellowship Programme: a partnership between the African Academy of Sciences and the Royal Society funded by the UK Government's Global Challenges Research Fund, to C.H.T.; and NSF grants 1565046 and 1661510, to C.M.

Author contributions A.L.P., C.H.T. and C.M. conceived the study, processed the species and climate data, performed the analysis and wrote the manuscript. All authors read and approved the final manuscript.

Competing interests The authors declare no competing interests.

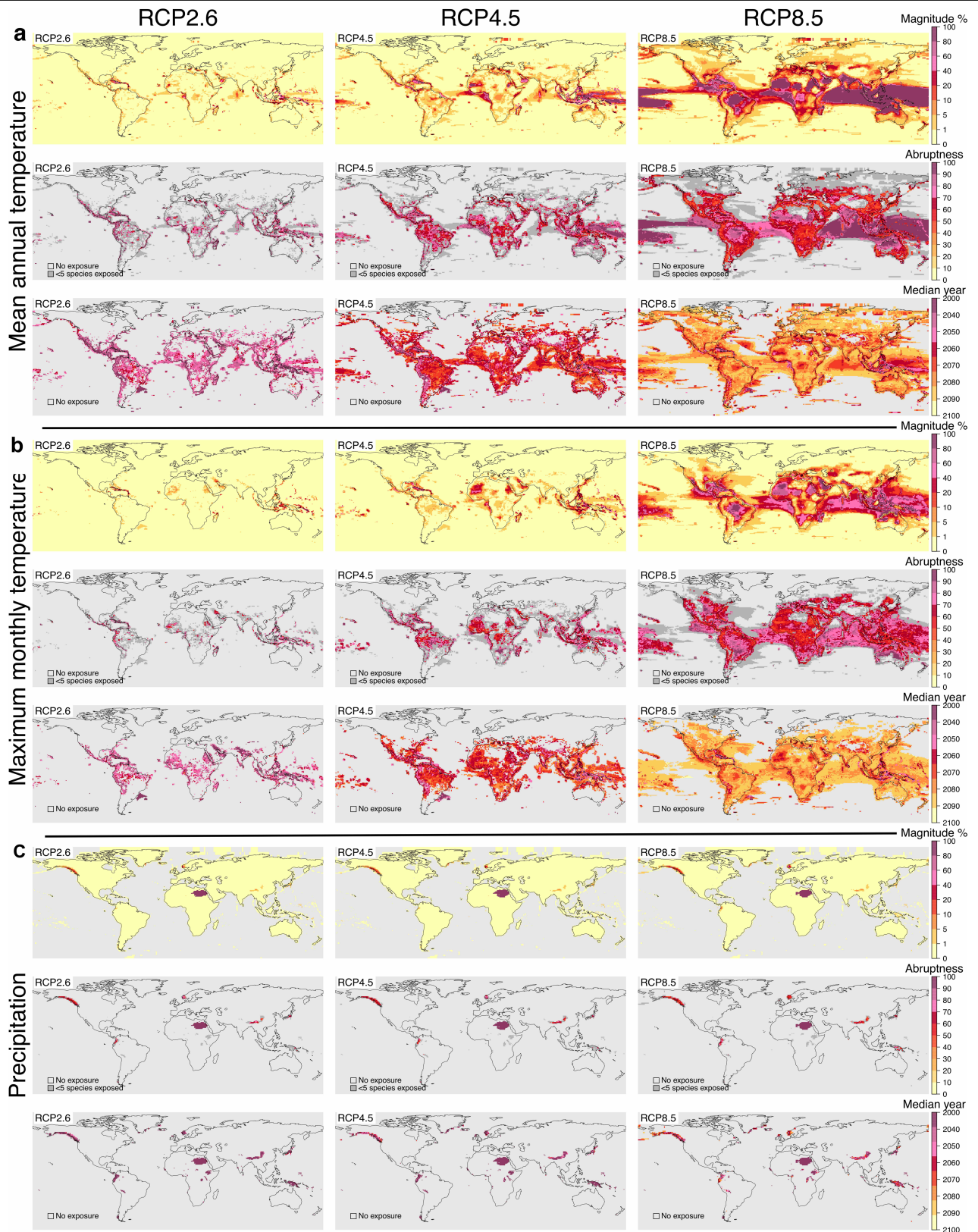
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2189-9>.

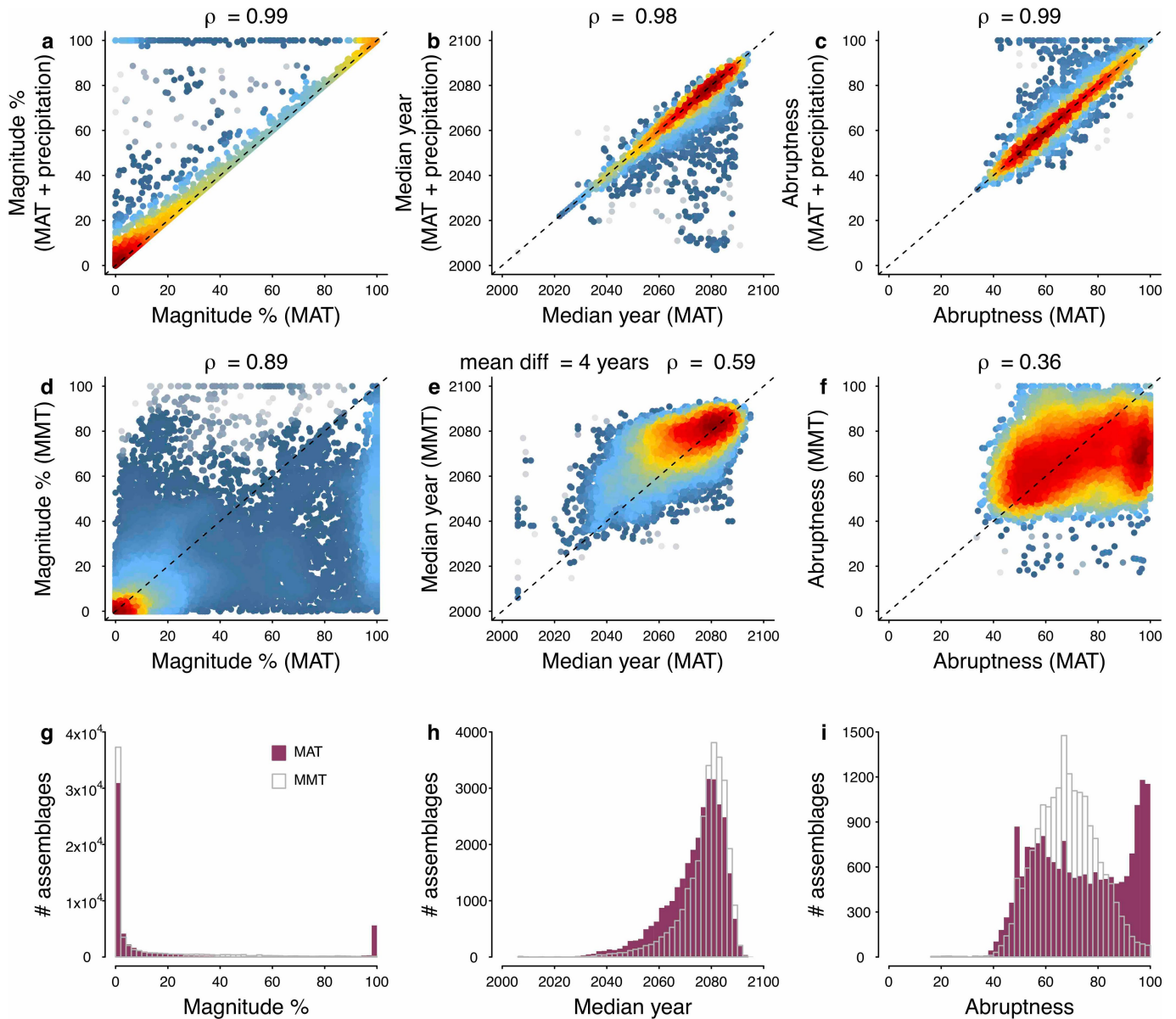
Correspondence and requests for materials should be addressed to A.L.P.

Peer review information Nature thanks Joanne Bennett, Anthony Richardson, Jennifer Sunday and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

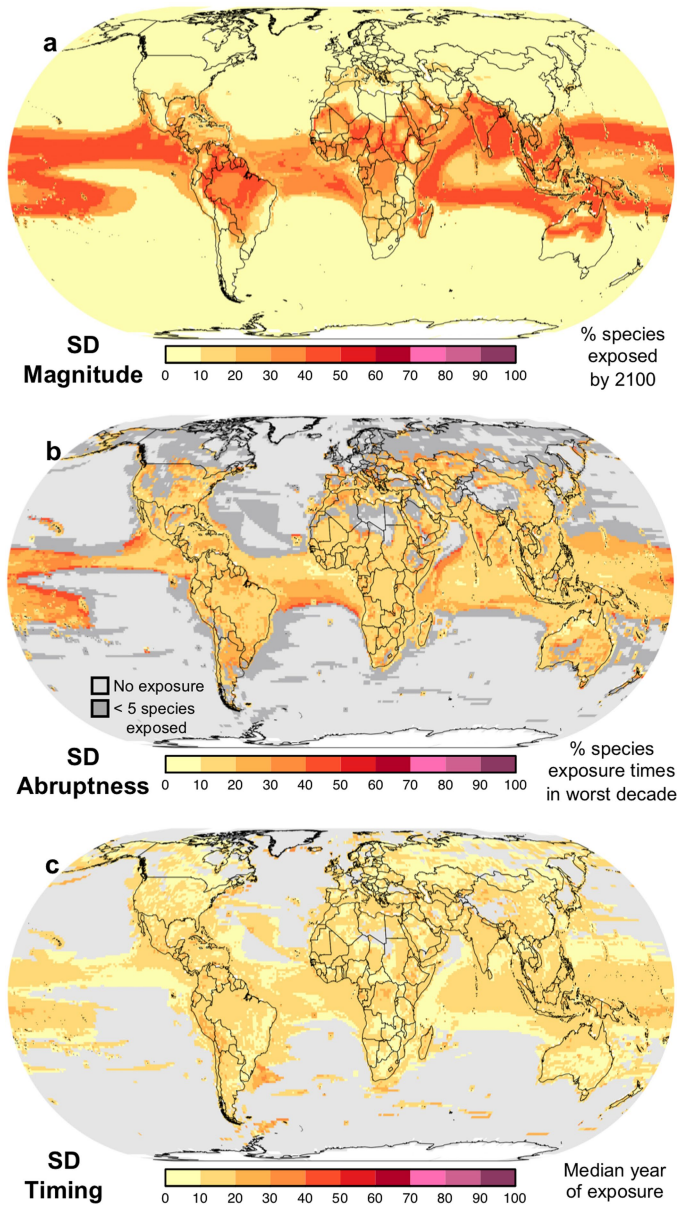


Extended Data Fig. 1 | Spatial distribution of the magnitude, abruptness and timing of assemblage exposure for alternative climate variables. a–c, Shown is the median value across 22 CMIP5 climate models for MAT (a), MMT (b) and precipitation (c) under RCP 2.6, RCP 4.5 and RCP 8.5.

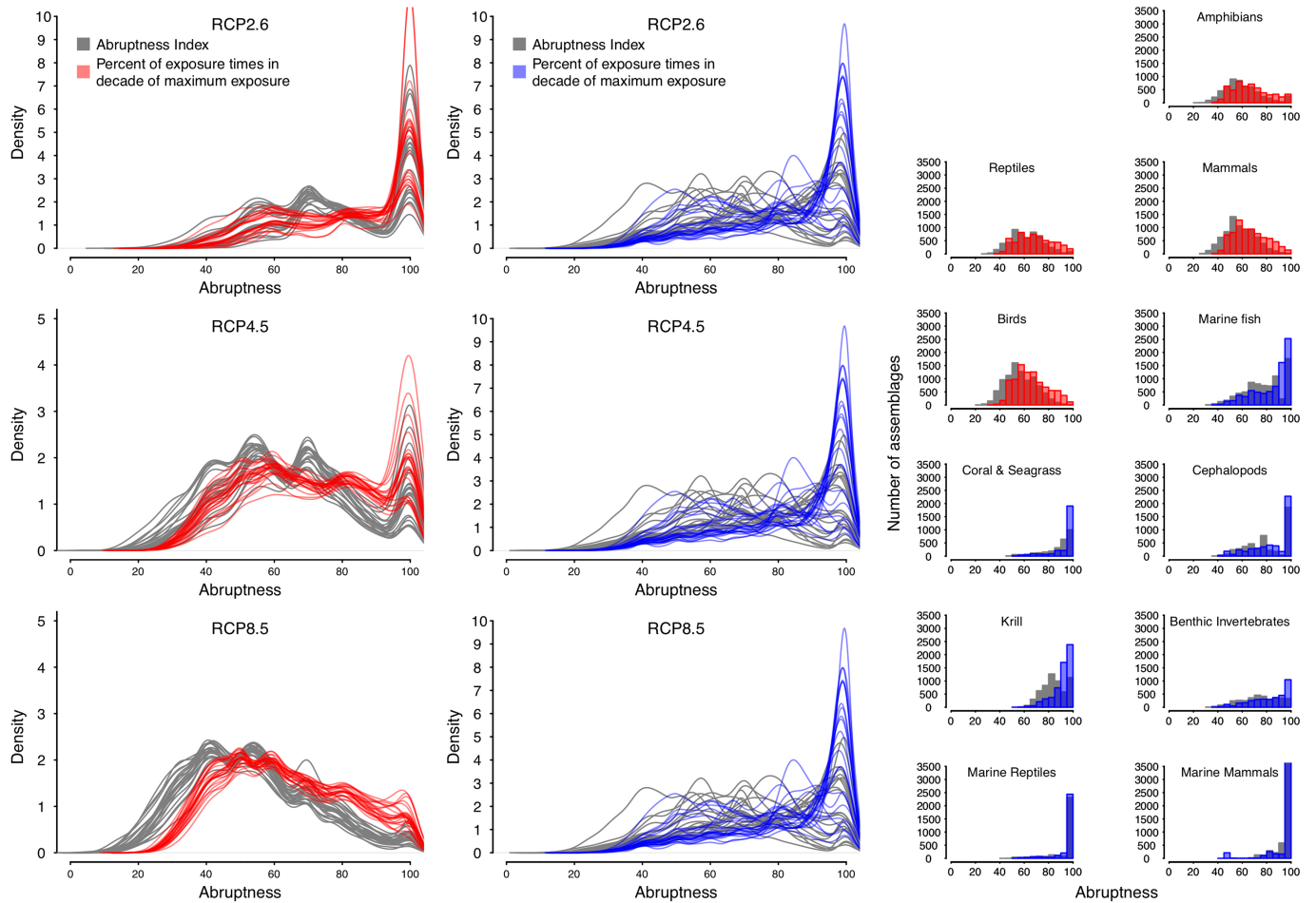


Extended Data Fig. 2 | Comparing the magnitude, timing and abruptness of assemblage exposure across alternative climate variables. **a–c**, Patterns of exposure to both MAT and precipitation combined are very similar to patterns of exposure to MAT only, highlighting the importance of changes in temperature in driving exposure. **d–i**, Patterns of exposure to unprecedented temperatures show both similarities and differences depending on whether temperature is quantified using MAT or MMT. More species are exposed and exposure occurs earlier for MAT compared with MMT, but spatial variation in the magnitude (**d, g**) and timing (**e, h**) of exposure are strongly correlated

between temperature variables. Variation in the abruptness of assemblage exposure is less strongly correlated between MAT and MMT (**f**), but both variables confirm the abruptness of projected exposure (**i**). Values are the median across 22 CMIP5 climate models under RCP 8.5, with hotter colours indicating a higher density of points. Points falling along the dashed 1:1 line indicate a perfect correspondence between metrics. The correlation between metrics (Spearman's ρ), and the mean difference in the timing of exposure (years), is shown.

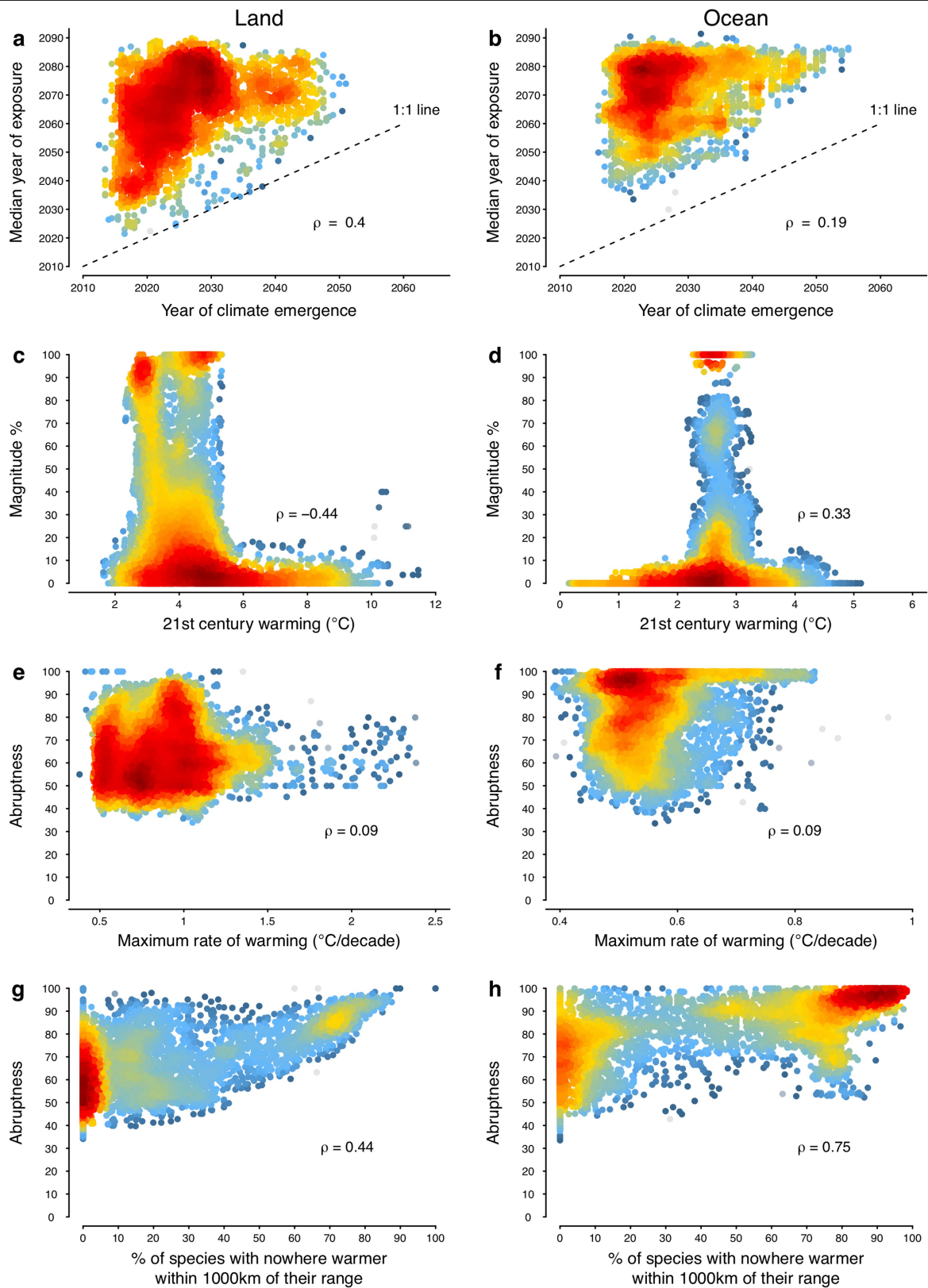


Extended Data Fig. 3 | Uncertainty in species local exposure metrics across 22 CMIP5 climate models under RCP 8.5. Uncertainty (standard deviation, SD) in the magnitude of exposure is greatest around the boundaries of the tropics, with little geographic variation in uncertainty in timing or abruptness.



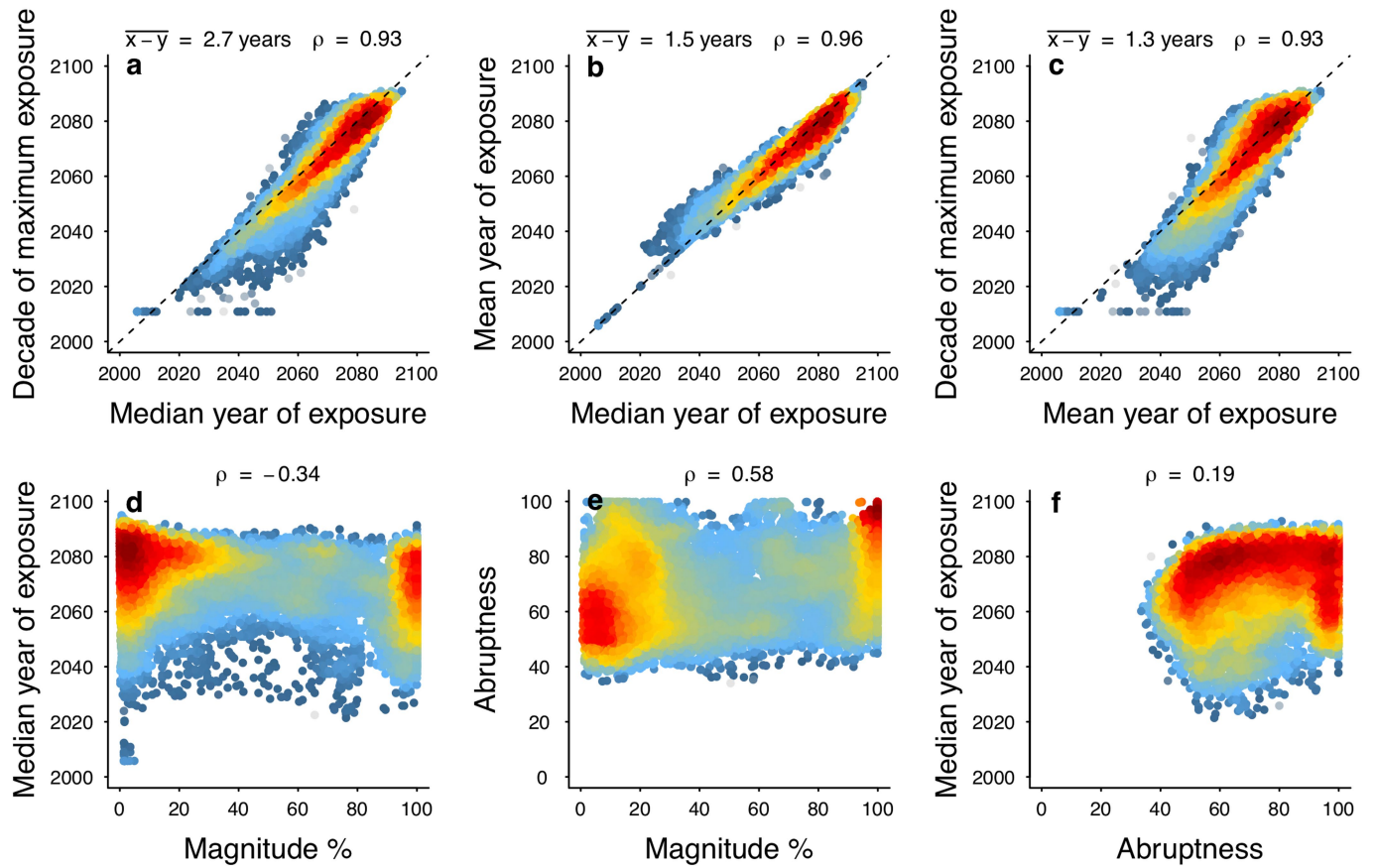
Extended Data Fig. 4 | Abruptness of horizon profiles. Density plots (left) show the distribution of abruptness values for different CMIP5 climate models ($n = 22$, lines) and RCPs on land (red) and in the ocean (blue). Histograms (right) show the median abruptness across climate models under RCP8.5 for each group of organisms. Abruptness is calculated as the percentage of exposure times occurring within the decadal window of maximum exposure (colours).

Abruptness is also shown for an alternative metric based on the Shannon entropy index (grey) with values scaled between 0 and 100, indicating the most gradual and the most abrupt distribution of exposure times possible for a given assemblage, respectively. Exposure is consistently abrupt across climate models, RCP scenarios, metrics and organism groups.



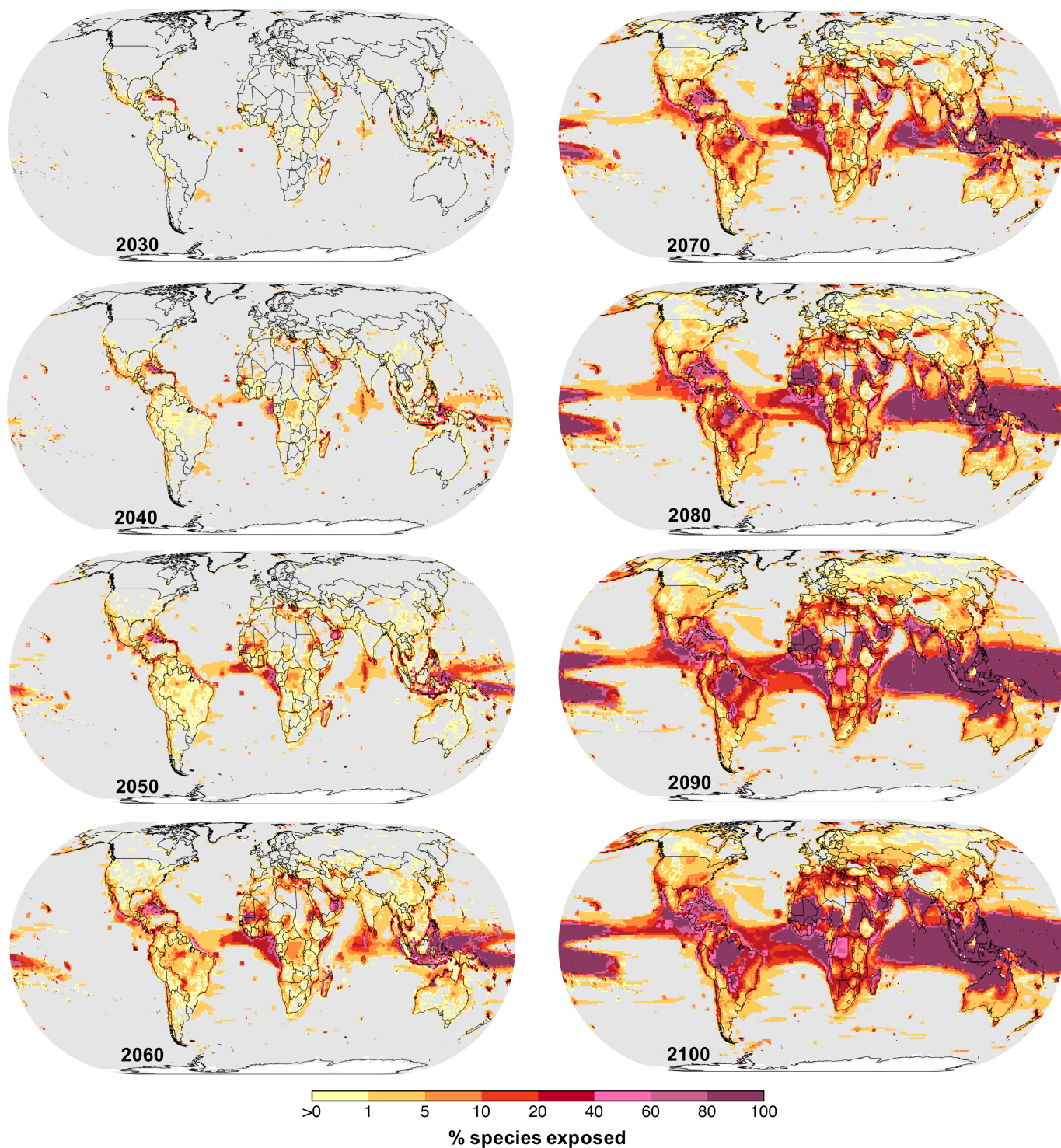
Extended Data Fig. 5 | Predicting the timing, magnitude and abruptness of local species exposure. **a–h**, On land (left) and in the ocean (right) the median timing of exposure (**a**, **b**) is weakly correlated (Spearman's ρ) with the timing of local climate emergence. The magnitude of exposure (**c**, **d**) is weakly correlated with the magnitude of warming between the start (2000–2020) and the end (2090–2100) of the twenty-first century. The abruptness of exposure

(percentage of local species exposure times that occur in the decade of maximum exposure) is only partly correlated with the maximum rate of warming (maximum difference in mean temperature between successive decades) (**e**, **f**) or the percentage of species with nowhere warmer within 1,000 km of their range (**g**, **h**). Values are the median across 22 CMIP5 climate models under RCP 8.5. Hotter colours indicate a higher density of points.



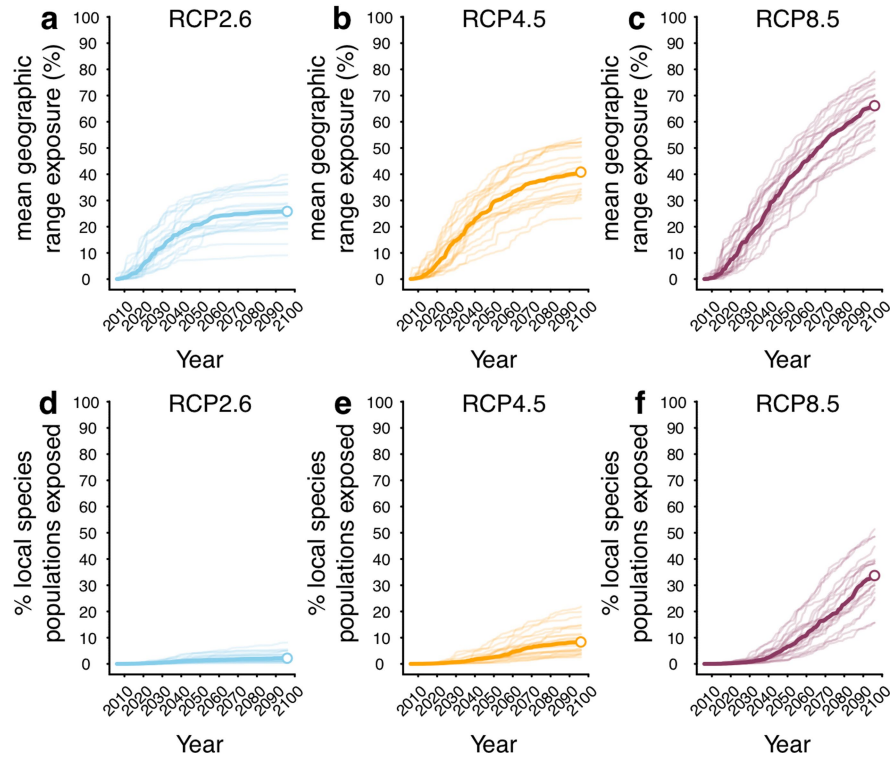
Extended Data Fig. 6 | The different dimensions of climate risk to species assemblages. **a–c**, Bivariate plots showing the strong correlation among alternative metrics for the timing of local assemblage exposure: the median year of local species exposure, the mean year of local species exposure and the mid-point of the decadal window of worst (that is, maximum) local species exposure. **d–f**, Bivariate plots showing the weaker correlation between the

magnitude, abruptness and timing of exposure across assemblages. Values are the median across 22 CMIP5 climate models under RCP 8.5, with hotter colours indicating a higher density of points. In **a–c**, points falling along the dashed 1:1 line indicate a perfect correspondence between metrics. The correlation between metrics (Spearman's ρ) is shown, as well as (for **a–c**) the mean difference in the timing of exposure (years).



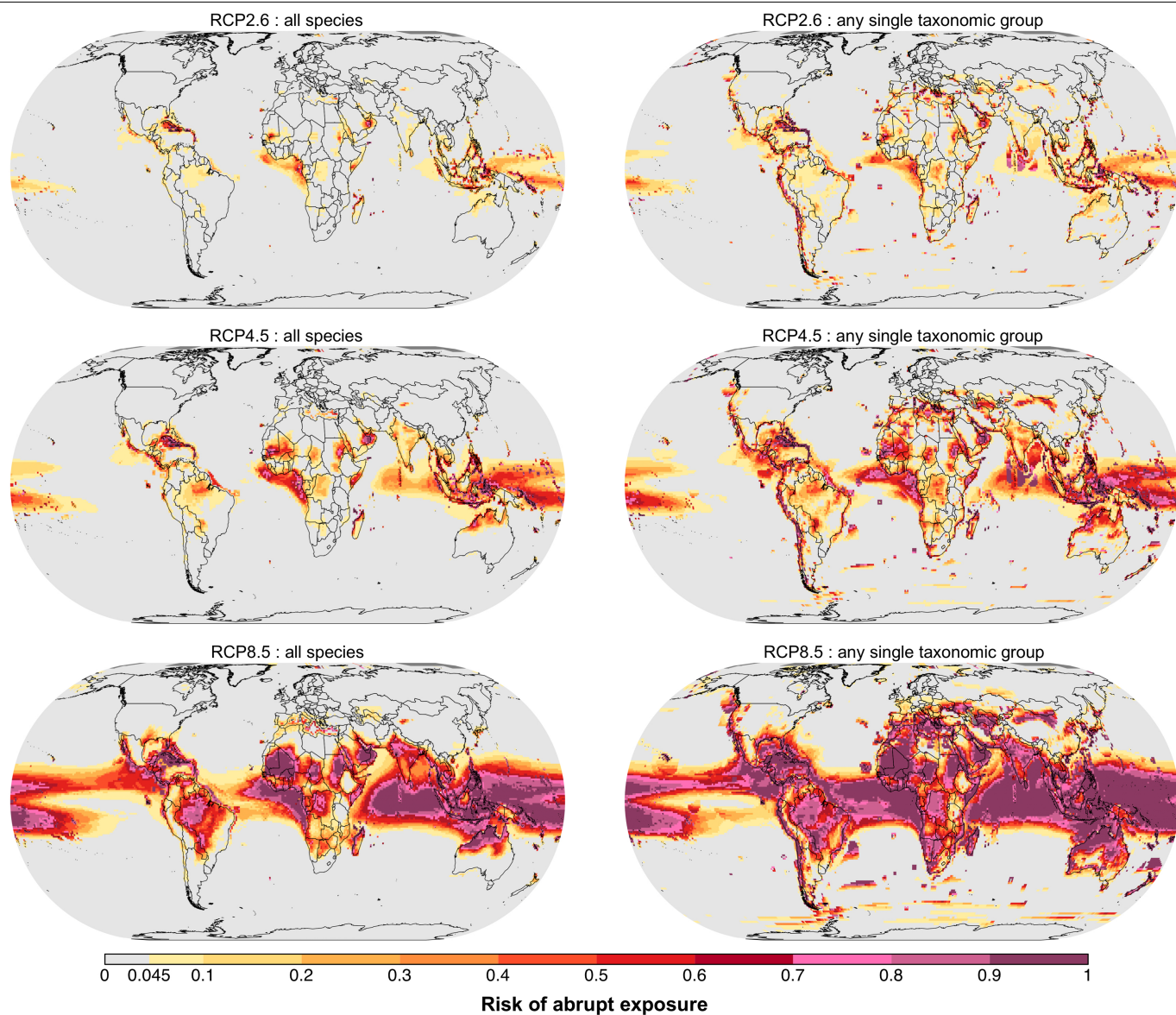
Extended Data Fig. 7 | Accumulation of exposure to unprecedented temperatures at decadal time snapshots from 2030 to 2100. Light grey indicates zero local species exposure. Maps show the median across 22 CMIP5

climate models under RCP 8.5, highlighting the immediate onset of exposure in the tropics that spreads to higher latitudes later in the century.



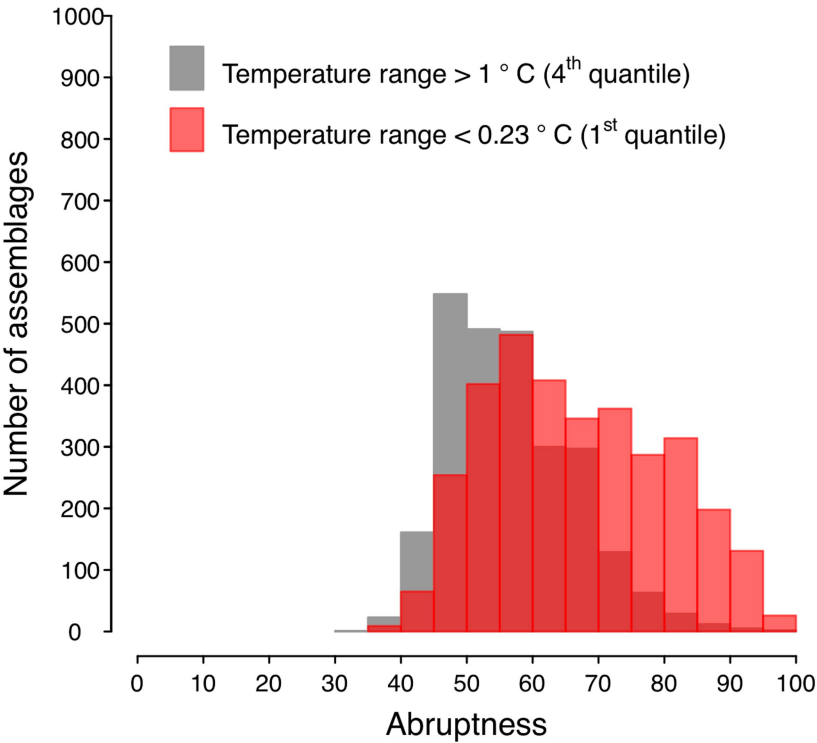
Extended Data Fig. 8 | The global biodiversity horizon profile. **a–d**, The cumulative exposure to unprecedented temperatures of all local species populations (that is, species X site aggregated across all sites) increases smoothly over time at the global scale. Global horizon profiles are shown when species are weighted by the inverse of their geographic range size (equivalent

to the mean percentage of the geographic range exposed) (**a, b**) or are given equivalent weighting (**d–f**). In **d–f**, dynamics are dominated by species with many local populations (that is, large geographic ranges). Variability in exposure across 22 climate models (thin lines) is shown for each RCP scenario (median, thick line).



Extended Data Fig. 9 | The global distribution in the risk of high-magnitude and abrupt assemblage exposure events under different representative concentration pathways. Maps show the probability of abrupt exposure calculated across 22 CMIP5 climate models. The risk of abrupt exposure was calculated on the basis of all species in an assemblage (left column) and for each

group of organisms separately (right column). The maps highlight the greater risk of abrupt exposure events under intermediate (RCP 4.5) and especially under high (RCP 8.5) emission pathways, and when considering taxonomic groups separately.



Extended Data Fig. 10 | Abruptness of horizon profiles for terrestrial vertebrates in 100-km grid cells with low or high spatial temperature heterogeneity. Red, low heterogeneity; grey, high heterogeneity. Abruptness is calculated as the percentage of species exposure times in the decade of maximum exposure. Temperature heterogeneity is the range in temperatures

at 1-km resolution within each 100-km cell. Assemblages with abrupt exposure have lower temperature heterogeneity, which suggests that quantifying species niches at finer spatial resolutions is unlikely to alter the abrupt nature of assemblage exposure dynamics.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- | | | |
|-------------------------------------|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A description of all covariates tested |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> | For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> | Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No software was used in the data collection process

Data analysis

Data analysis was performed in R v 3.6.1. Computer code used in the analysis is available on request from the authors. Code and results data to make figures 2-4 is available at figshare (10.6084/m9.figshare.11814633).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All datasets used are publicly available. We used expert verified range maps for 30,652 species from the International Union for Conservation of Nature (<https://www.iucnredlist.org/resources/spatial-data-download>) and BirdLife International (<http://datazone.birdlife.org/species/requestdis>) including: birds, mammals, reptiles, amphibians, marine fish, benthic marine invertebrates, and habitat forming corals and seagrasses. Climate change projections for RCPs 8.5, 4.5, and 2.6 for the Coupled Model Intercomparison Project 5 (CMIP5) are available from <https://esgf-node.llnl.gov/search/cmip5/>. Results data to make Figures 2-4 is available at figshare (10.6084/m9.figshare.11814633).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Future climate projections from earth system models are combined with information on species geographic distributions (n=30,652 species) to estimate species realised thermal niche limits and project the timing of future exposure to conditions beyond their niche.
Research sample	We used expert verified range maps for 30,652 species from the International Union for Conservation of Nature (https://www.iucnredlist.org/resources/spatial-data-download) and BirdLife International (http://datazone.birdlife.org/species/requestdis), including; birds, mammals, reptiles, amphibians, marine fish, benthic marine invertebrates, and habitat forming corals and seagrasses. This sample reflects availability of geographic range data for each organisms group globally.
Sampling strategy	The sample size reflects availability of geographic range data for each organisms group. These sample sizes make up the majority of species in each major taxonomic group included in the study.
Data collection	All data used here is already published and was downloaded from public data portals
Timing and spatial scale	The data on species distributions we use represents more than a century of collecting efforts and observations by scientists, naturalists and the public. The climate data we use is generated by simulations from computer models of the earth system. Both kinds of data are accurate to ~100km resolution and are available globally.
Data exclusions	We excluded marine species restricted to depths >200m to focus on the effects of sea surface temperatures on shallow water species. To prevent estimates of maximum temperature being inflated by either extreme outliers in the temperature time series or from the overestimation of species ranges we excluded outlier temperature values within each grid cell, defined as those more than three standard deviations from the mean. Once we had selected the maximum temperature for each cell, we excluded outlier temperature values across each species range, defined as those more than three standard deviations above the mean range value.
Reproducibility	This is not an experimental study so experimental replication was not attempted. All data used in our analysis is publicly available. The results can be reproduced using the publicly available data and the analysis code is available on request from the authors. Code and results data to reproduce Figures 2-4 is available at figshare (10.6084/m9.figshare.11814633).
Randomization	No randomization was required. Our study was not experimental, but based on observed biodiversity and climate patterns.
Blinding	Our study was not experimental and so blinding is not relevant
Did the study involve field work?	<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Extant timetrees are consistent with a myriad of diversification histories

<https://doi.org/10.1038/s41586-020-2176-1>

Stilianos Louca^{1,2}✉ & Matthew W. Pennell^{3,4}✉

Received: 14 September 2019

Accepted: 10 March 2020

Published online: 15 April 2020

 Check for updates

Time-calibrated phylogenies of extant species (referred to here as ‘extant timetrees’) are widely used for estimating diversification dynamics¹. However, there has been considerable debate surrounding the reliability of these inferences^{2–5} and, to date, this critical question remains unresolved. Here we clarify the precise information that can be extracted from extant timetrees under the generalized birth–death model, which underlies most existing methods of estimation. We prove that, for any diversification scenario, there exists an infinite number of alternative diversification scenarios that are equally likely to have generated any given extant timetree. These ‘congruent’ scenarios cannot possibly be distinguished using extant timetrees alone, even in the presence of infinite data. Importantly, congruent diversification scenarios can exhibit markedly different and yet similarly plausible dynamics, which suggests that many previous studies may have over-interpreted phylogenetic evidence. We introduce identifiable and easily interpretable variables that contain all available information about past diversification dynamics, and demonstrate that these can be estimated from extant timetrees. We suggest that measuring and modelling these identifiable variables offers a more robust way to study historical diversification dynamics. Our findings also make it clear that palaeontological data will continue to be crucial for answering some macroevolutionary questions.

A central challenge in evolutionary biology is to reconstruct rates of speciation and extinction over time⁵. Unfortunately, the majority of taxa that have ever lived have not left much trace in the fossil record, and the primary source of information on their past diversification dynamics therefore comes from extant timetrees. Many methods have been developed for extracting this information; most methods fit variants of a birth–death process^{1,6}. Despite the popularity of these methods, which collectively have been used in thousands of studies^{7–9}, their reliability has been called into question by comparisons with fossil-based estimates^{1,3,5,6,10}. The reasoning behind these critiques is that there may be insufficient information in extant timetrees to fully reconstruct historical diversification dynamics. However, this critical issue has remained unresolved; it is unknown precisely what information on speciation and extinction rates is contained in extant timetrees.

Here we present a definite answer to this question for the general stochastic birth–death process with homogeneous (that is, lineage-independent) rates, in which speciation (‘birth’) rates (λ) and extinction (‘death’) rates (μ) can vary over time, that underlies the majority of existing methods for reconstructing diversification dynamics from phylogenies¹. We mathematically show that, for any given candidate birth–death model, there exists an infinite number of alternative birth–death models that can explain any extant timetree equally as well as can the candidate model. These alternative models may appear to be similarly plausible and yet exhibit markedly different features, such as

different trends through time in both λ and μ . This severe ambiguity persists for arbitrarily large trees and cannot be resolved even with an infinite amount of data; it is thus impossible to design asymptotically consistent estimators for λ and μ . Using simulated and real timetrees as examples, we demonstrate how failing to recognize this issue can seriously mislead our inferences about past diversification dynamics. We present appropriately modified variables that are asymptotically identifiable and that contain all available information on historical diversification dynamics.

Lineages through time

An important feature of extant timetrees is the lineages-through-time curve (LTT), which counts the number of lineages at each time in the past that are represented by at least one sampled extant descending species in the tree. The likelihood of a tree under a given birth–death model, the LLT of the tree and the LTT that would be expected under the model are linked as follows. Any given combination of (potentially time-dependent) speciation and extinction rates (λ and μ , respectively) and the probability that an extant species will be included in the tree (‘sampling fraction’) (ρ) can be used to define a deterministic diversification process, in which the number of lineages through time no longer varies stochastically but instead according to a set of differential equations^{6,11} (Supplementary Information section S.1). Given a number of extant sampled species (M_0), the LTT predicted by these

¹Department of Biology, University of Oregon, Eugene, OR, USA. ²Institute of Ecology and Evolution, University of Oregon, Eugene, OR, USA. ³Biodiversity Research Centre, University of British Columbia, Vancouver, British Columbia, Canada. ⁴Department of Zoology, University of British Columbia, Vancouver, British Columbia, Canada. ✉e-mail: louca.research@gmail.com; pennell@zoology.ubc.ca

differential equations (the deterministic LTT (dLTT)) corresponds to the LTT that is expected for trees generated by the original stochastic model¹¹. The likelihood of a tree under a given birth–death model can be written purely in terms of the LTT of the tree and the dLTT of the model (see Supplementary Information section S.1.2 and ref.¹² for derivations). This means that any two models with the same dLTT (conditioned on M_0) yield identical likelihoods for the tree. We therefore call two models ‘congruent’ if they have the same dLTT for any given M_0 . Any two models are either congruent or non-congruent, regardless of any particular data considered (Supplementary Information section S.1). The probability distribution of tree sizes generated by a model, when conditioned on the age of the stem or crown, is identical among congruent models (Supplementary Information section S.1.7). Hence, congruent models have equal probabilities of generating any given timetree, analogous to how congruent geometric objects exhibit similar properties (discussion in Supplementary Information section S.1.8). Although the mathematical relationship between the dLTT of a model and its likelihood has been known¹², its implications for macro-evolutionary inference have remained unexamined and—as we show below—severely underestimated.

The breadth of congruent model sets

When seen as a random variable, extant timetrees have the same probability distribution under any two congruent models. Therefore, in the absence of further information, congruent models cannot possibly be distinguished solely on the basis of extant timetrees—neither through the likelihood nor any other test statistic. For any birth–death model, this leads to four important unresolved questions: how many alternative congruent models there are, how different these congruent models are from one another, how many of these congruent models correspond to plausible scenarios, and how these scenarios can be explored. To answer these questions, we present an alternative method for recognizing congruent models (full details are provided in Supplementary Information section S.1.1). Given a number of sampled species (M_0), the dLTT of a model is fully determined by its relative slope (hereafter, pulled speciation rate), denoted by $\lambda_p = -M^{-1}dM/d\tau$ (in which M is the dLTT, τ is time before present (or age) and p is a label (for ‘pulled’)). It can be shown that $\lambda_p = \lambda P$, in which $P(\tau)$ is the probability that a lineage extant at age τ survives until the present day and is included in the timetree. In the absence of extinction ($\mu = 0$) and under complete species sampling ($\rho = 1$), λ_p is identical to λ ; however, in the presence of extinction λ_p is pulled downwards relative to λ at older ages, whereas under incomplete sampling λ_p is pulled downwards relative to λ near the present. Because the dLTT of a model is fully determined by λ_p and vice versa, two models are congruent if and only if they have the same λ_p at all ages. In a similar way, it can be shown that two models are congruent if and only if they have the same product $\rho\lambda_0$ (in which $\lambda_0 = \lambda(0)$) and the same ‘pulled diversification rate’¹³, which is another composite variable and is defined as

$$r_p = \lambda - \mu + \frac{1}{\lambda} \frac{d\lambda}{d\tau} \quad (1)$$

The r_p is equal to the net diversification rate ($r = \lambda - \mu$) whenever λ is constant in time ($d\lambda/d\tau = 0$), but differs from r when λ varies with time.

We are now ready to examine the breadth of congruent model sets. We begin with a model with speciation rate $\lambda > 0$, extinction rate $\mu \geq 0$ and sampling fraction $\rho \in (0, 1]$. If we denote $\eta_0 = \rho\lambda_0$, then for any alternative chosen extinction rate function $\mu^* \geq 0$ and any alternative assumed sampling fraction $\rho^* \in (0, 1]$, there exists a speciation rate function $\lambda^* > 0$ such that the alternative model (λ^*, μ^* and ρ^*) is congruent to the original model (λ, μ and ρ). In other words, regardless of the chosen μ^* and ρ^* , we can find a hypothetical λ^* that satisfies $\lambda^* - \mu^* + (1/\lambda^*)d\lambda^*/d\tau = r_p$ and

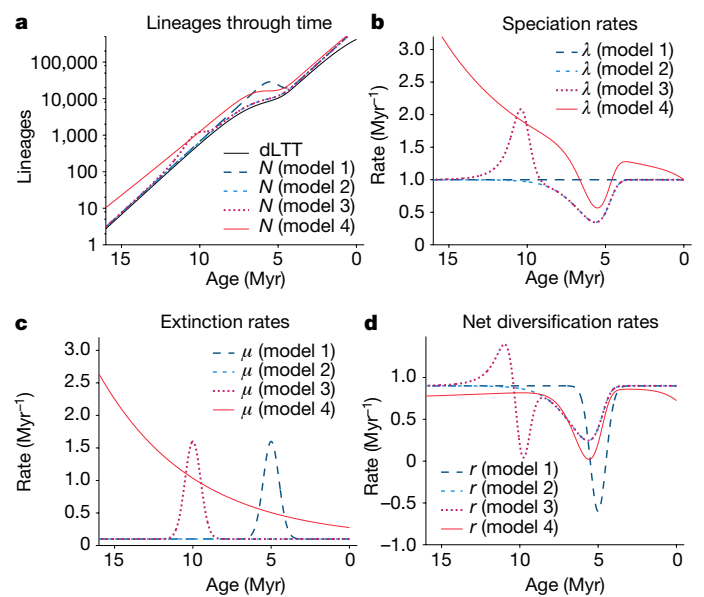


Fig. 1 | Illustration of congruent birth–death processes (simulations).

Example of four hypothetical congruent—yet markedly different—birth–death models. All models exhibit the same dLTT, and would yield the same likelihood for any given extant timetree. **a**, dLTT and deterministic diversities (N) predicted by the models, plotted over age (time before present).

b–d, Speciation rates (λ) (**b**), extinction rates (μ) (**c**) and net diversification rates ($r = \lambda - \mu$) (**d**) of the models. Myr, million years. For additional examples, see Extended Data Fig. 4.

$\rho^*\lambda^*(0) = \eta_0$. Indeed, to construct such a λ^* one merely needs to solve the following differential equation:

$$\frac{d\lambda^*}{d\tau} = \lambda^* \cdot (r_p - \lambda^* + \mu^*) \quad (2)$$

with initial condition $\lambda^*(0) = \eta_0/\rho^*$ (solution in Supplementary Information section S.1.4). The above observation implies that—starting from almost any birth–death model—we can generate an infinite number of alternative congruent models simply by modifying the extinction rate (μ) and/or the assumed sampling fraction (ρ). Alternatively, congruent models can be constructed by assuming various ratios of μ/λ (Supplementary Information section S.1.5). This set of congruent models (hereafter, the congruence class) is thus infinitely large. The congruence class can have an arbitrary number of dimensions (depending on restrictions imposed a priori on λ^* and μ^*), as μ^* could depend on an arbitrarily high number of free parameters.

As an illustration of these principles, the simulations in Fig. 1 show four markedly distinct and yet congruent models (pulled rates are shown in Extended Data Fig. 1). The first scenario exhibits a constant λ and a temporary spike in μ (that is, a mass extinction event), the second scenario exhibits a constant μ and a temporary drop in λ around the same time, the third scenario exhibits a mass extinction event at a completely different time and a fluctuating λ , and the fourth scenario exhibits an exponentially decaying μ and a fluctuating λ . These congruent scenarios were obtained simply by assuming alternative extinction rates, and a myriad of other congruent scenarios exist. Figure 2 shows a model with exponentially varying speciation and extinction rates, $\lambda = \alpha e^{\beta\tau}$ and $\mu = \gamma e^{\delta\tau}$, with α, β, γ and δ fitted to a timetree of 79,874 extant seed plant species¹⁴ via maximum-likelihood methods. Simply by modifying the coefficient δ and choosing λ according to equation (2), one can obtain a similarly complex scenario with opposite trends over time (Fig. 2b). Similar examples can also be generated using more realistic speciation and extinction rates, three of which can be seen in Extended

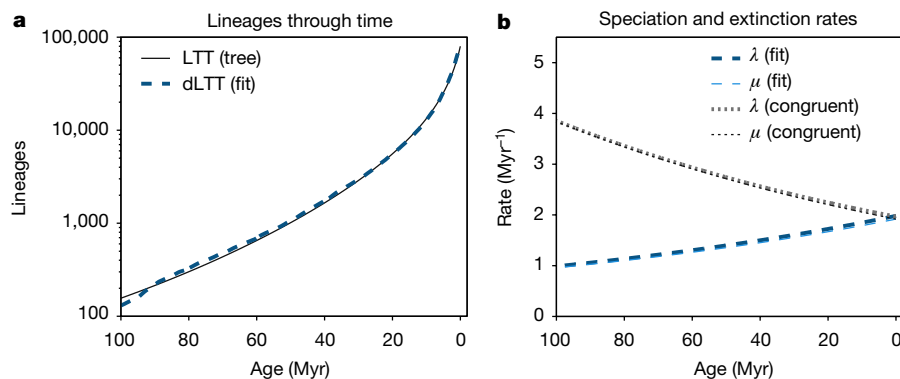


Fig. 2 | Illustration of congruent birth–death processes (real data). Birth–death model with exponentially varying λ and μ , fitted to an extant timetree of 79,874 seed plant species¹⁴ over the past 100 Myr, compared to a congruent model obtained by simply modifying the exponential coefficient of μ . **a**, LTT of the tree, compared to the dLTT predicted by the two models. **b**, Speciation

rates (λ) and extinction rates (μ) of the two models. In each model, μ is almost identical to λ . The two models cannot possibly be distinguished using extant timetrees alone. For additional examples, see Extended Data Fig. 2 and Supplementary Information sections S.10 and S.11.

Data Fig. 2 (based on data from ref.¹⁵), Extended Data Fig. 3 (based on data from ref.¹⁰) and Extended Data Fig. 4.

Such ambiguities have previously been observed in special cases^{11,16}. For example, a previous study¹¹ recognized that a variable λ and constant μ can be exchanged for a constant λ and a variable μ to produce the same dLTT. Other work on constant-rate birth–death models has revealed that alternative combinations of time-independent λ , μ and ρ can yield the same likelihood for a tree^{17,18}. Our work not only unifies these previous findings (which are all special cases of our general theory), but in fact reveals that vast (infinite-dimensional) expanses of model space are fundamentally indistinguishable even if ρ is known or all extant species have been sampled.

Implications

To estimate λ and μ , previous phylogenetic studies have imposed largely arbitrary constraints. For example, many studies assume that λ or μ vary exponentially through time¹⁹. However, this functional form is rarely justified biologically, and alternative functional forms of comparable simplicity and shape can be envisioned. Normally one expects that, with sufficient data, fitting any of these forms will lead to qualitatively similar trends and shapes. This expectation simply does not hold here, because the best-fitting representative within a given model set will generally only be the one closest to the congruence class of the true process, rather than closest to the true process itself (Fig. 3). Consequently, fitting alternative functional forms can result in markedly different inferences with alternative trends in λ and μ , even if each functional form used is in principle adequate for approximating the true historical λ and μ (examples are shown in Extended Data Fig. 5, Supplementary Information section S.10). This conclusion applies to almost any model set used in practice, including models in which λ and μ change at discrete time points²⁰. Because any given true diversification history (even a relatively simple one) is unlikely to exactly match the particular functional form considered, fitting the latter may not even approximately yield the true diversification history. The existence of congruent scenarios can thus seriously alter macroevolutionary conclusions—for example, when assessing the influence of environmental factors on diversification dynamics (example shown in Supplementary Information section S.4, and further discussion in Supplementary Information section S.5). Our findings thus shed doubt over previous work on diversification dynamics that is based solely on extant timetrees, including some of the conclusions from work that we have coauthored^{9,13}. Previous studies have underestimated this issue because they typically consider only a limited set of candidate models at a time, both when analysing real datasets and when assessing

parameter identifiability via simulations; as a result, previous studies have been (un)lucky enough to not compare two models in the same congruence class (see Supplementary Information sections S.3 and S.7 for reasoning). We stress that common model selection methods that are based on parsimony or ‘Occam’s razor’ (such as the Akaike information criterion²¹) generally cannot resolve these issues (Extended Data Fig. 6, details in Supplementary Information sections S.2 and S.10).

Ways forward

Our findings are analogous to classic results from coalescent theory in population genetics, in which many alternative models can give rise to the

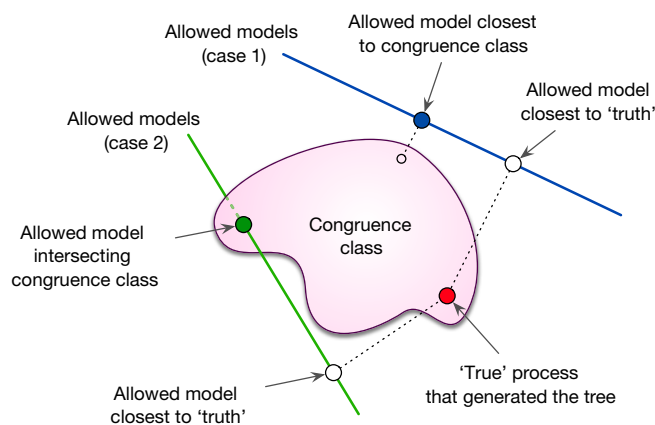


Fig. 3 | Conceptual implications. Conceptual illustration of the limited identifiability of a diversification process, assumed to be adequately described by some unknown birth–death model (red circle; hereafter ‘true process’). The congruence class of the true process is shown as a sub-space comprising a continuum of alternative models (pink area). In practice, maximum-likelihood model selection is performed among a parameterized low-dimensional set of allowed models, the precise nature of which can vary from case to case (for example, depending on assumed functional forms for λ and μ , or the number of allowed rate shifts²⁰). The two continuous lines shown here represent two alternative cases of allowed model sets (for example, considered in two alternative studies), from within each the model closest to the truth is (ideally) sought. However, in each case likelihood-based model selection will converge towards the allowed model closest to the congruence class (blue and green filled circles), which in general is not the allowed model that is actually closest to the true process (white circles). This identifiability issue persists even for infinitely large datasets.

same drift process as the idealized Wright–Fisher model^{22,23}. This realization was particularly important for the field: it focused the attention of researchers on the dynamics of the effective population size, an identifiable parameter, rather than on actual (but non-identifiable) historical demography. Similarly, congruent birth–death models can be defined in terms of λ_p or—equivalently—in terms of r_p and $\rho\lambda_p$, all of which are identifiable provided sufficient data¹³. Each congruence class contains exactly one model with $\mu = 0$ and $\rho = 1$, which is also the model in which $\lambda = \lambda_p$; hence, the pulled speciation rate can be interpreted as the speciation rate that generates the dLTT of the congruence class in the absence of extinctions and under complete species sampling. In other words, λ_p can be seen as the ‘effective’ speciation rate that fully explains the shape of the LTT of the tree. Similarly, each congruence class contains models with time-independent λ , and for these models $r_p = r$; therefore, the pulled diversification rate can be interpreted as the effective net diversification rate if λ was time-independent.

Fossil data could help to resolve the ambiguities highlighted here^{24,25}, and biological knowledge could, in principle, also help to reduce ambiguities. For example, if ρ and μ are somehow known from other sources, the congruence class collapses to a unique diversification scenario (Extended Data Fig. 7). Nevertheless, for many taxa the fossil record remains scarce and ambiguous, and our general understanding of what constitutes a plausible diversification scenario is poorly developed. Rather than attempting to estimate λ and μ , one can estimate λ_p , r_p and $\rho\lambda_p$ (and λ_p , if ρ is known)—for example, by using likelihood methods²⁶ (Extended Data Fig. 8, Supplementary Information section S.9). Previous work¹³ has shown that r_p can indeed yield insight into diversification dynamics and help to detect major transitions over time (Supplementary Information section S.8), as changes in r_p necessarily imply changes in λ and/or μ . Through λ_p , it also becomes possible to simulate and analyse diversification models with substantially simplified mathematical tools, as any model is congruent to a model with speciation rate λ_p , zero extinction and complete species sampling²⁷. Reciprocally, many existing estimation tools can be used to estimate λ_p and r_p by constraining μ to be zero or λ to be time-independent, respectively. Depending on the situation, other invariants of congruence classes may also have advantages, such as the ‘coalescent density’ introduced by ref. ¹², which permits an elegant description of the distribution of branching ages (see Supplementary Information section S.8 for further details).

Conclusions

Without further information or biologically well-justified constraints, in general extant timetrees alone cannot be used to reliably infer speciation rates (except for the present day), extinction rates or net diversification rates. Consequently, correlations between λ , μ or r and fluctuating environmental factors (such as temperature) also cannot be reliably inferred, neither when λ , μ or r are first estimated and then related to the environmental factors nor if λ , μ and r are expressed as parameterized functions of the environmental factors and then fitted to the timetree (Supplementary Information section S.5), because different parameterizations can lead to completely different inferences. Our findings could explain why diversification dynamics observed in the fossil record often contradict inferences based on phylogenetics^{1,3,5,6,10}, although other explanations have also been proposed^{28,29}. It is possible that similar major identifiability issues may also be hiding in other evolutionary reconstruction methods based on extant organisms alone, but this remains to be examined. On a more positive note, we have resolved a long-standing debate and precisely clarified what information can be extracted from extant timetrees alone, formulated in terms of easily interpretable and identifiable variables.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2176-1>.

- Morlon, H. Phylogenetic approaches for studying diversification. *Ecol. Lett.* **17**, 508–525 (2014).
- Quental, T. B. & Marshall, C. R. Extinction during evolutionary radiations: reconciling the fossil record with molecular phylogenies. *Evolution* **63**, 3158–3167 (2009).
- Quental, T. B. & Marshall, C. R. Diversity dynamics: molecular phylogenies need the fossil record. *Trends Ecol. Evol.* **25**, 434–441 (2010).
- Liow, L. H., Quental, T. B. & Marshall, C. R. When can decreasing diversification rates be detected with molecular phylogenies and the fossil record? *Syst. Biol.* **59**, 646–659 (2010).
- Marshall, C. R. Five palaeobiological laws needed to understand the evolution of the living biota. *Nat. Ecol. Evol.* **1**, 0165 (2017).
- Morlon, H., Parsons, T. L. & Plotkin, J. B. Reconciling molecular phylogenies with the fossil record. *Proc. Natl Acad. Sci. USA* **108**, 16327–16332 (2011).
- Rabosky, D. L. et al. An inverse latitudinal gradient in speciation rate for marine fishes. *Nature* **559**, 392–395 (2018).
- Condamine, F. L., Rolland, J. & Morlon, H. Assessing the causes of diversification slowdowns: temperature-dependent and diversity-dependent models receive equivalent support. *Ecol. Lett.* **22**, 1900–1912 (2019).
- Henao Diaz, L. F., Harmon, L. J., Sugawara, M. T. C., Miller, E. T. & Pennell, M. W. Macroevolutionary diversification rates show time dependency. *Proc. Natl Acad. Sci. USA* **116**, 7403–7408 (2019).
- Steehan, M. E. et al. Radiation of extant cetaceans driven by restructuring of the oceans. *Syst. Biol.* **58**, 573–585 (2009).
- Kubo, T. & Iwasa, Y. Inferring the rates of branching and extinction from molecular phylogenies. *Evolution* **49**, 694–704 (1995).
- Lambert, A. & Stadler, T. Birth–death models and coalescent point processes: the shape and probability of reconstructed phylogenies. *Theor. Popul. Biol.* **90**, 113–128 (2013).
- Louca, S. et al. Bacterial diversification through geological time. *Nat. Ecol. Evol.* **2**, 1458–1467 (2018).
- Smith, S. A. & Brown, J. W. Constructing a broadly inclusive seed plant phylogeny. *Am. J. Bot.* **105**, 302–314 (2018).
- Alroy, J. Colloquium paper: dynamics of origination and extinction in the marine fossil record. *Proc. Natl Acad. Sci. USA* **105** (Suppl 1), 11536–11542 (2008).
- Stadler, T. Simulating trees with a fixed number of extant species. *Syst. Biol.* **60**, 676–684 (2011).
- Stadler, T. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. *J. Theor. Biol.* **261**, 58–66 (2009).
- Stadler, T. & Steel, M. Swapping birth and death: symmetries and transformations in phylodynamic models. *Syst. Biol.* **68**, 852–858 (2019).
- Rabosky, D. L. Automatic detection of key innovations, rate shifts, and diversity-dependence on phylogenetic trees. *PLoS ONE* **9**, e89543 (2014).
- Stadler, T. Mammalian phylogeny reveals recent diversification rate shifts. *Proc. Natl Acad. Sci. USA* **108**, 6187–6192 (2011).
- Akaike, H. Likelihood of a model and information criteria. *J. Econom.* **16**, 3–14 (1981).
- Möhle, M. Robustness results for the coalescent. *J. Appl. Probab.* **35**, 438–447 (1998).
- Sjödin, P., Kaj, I., Krone, S., Lascoux, M. & Nordborg, M. On the meaning and existence of an effective population size. *Genetics* **169**, 1061–1070 (2005).
- Heath, T. A., Huelsenbeck, J. P. & Stadler, T. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proc. Natl Acad. Sci. USA* **111**, E2957–E2966 (2014).
- Stadler, T., Gavryushkina, A., Warnock, R. C. M., Drummond, A. J. & Heath, T. A. The fossilized birth–death model for the analysis of stratigraphic range data under different speciation modes. *J. Theor. Biol.* **447**, 41–55 (2018).
- Louca, S. & Doebeli, M. Efficient comparative phylogenetics on large trees. *Bioinformatics* **34**, 1053–1055 (2018).
- Louca, S. Simulating trees with millions of species. *Bioinformatics* <https://doi.org/10.1093/bioinformatics/btaa031> (2020).
- Rabosky, D. L. Heritability of extinction rates links diversification patterns in molecular phylogenies and fossils. *Syst. Biol.* **58**, 629–640 (2009).
- Silvestro, D., Warnock, R. C. M., Gavryushkina, A. & Stadler, T. Closing the gap between palaeontological and neontological speciation and extinction rate estimates. *Nat. Commun.* **9**, 5237 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Article

Methods

No statistical methods were used to predetermine sample size. Thorough mathematical derivations and computational details are provided in Supplementary Information sections S.1–S.11.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

No new data were generated for this manuscript. All phylogenetic datasets used as examples have previously been published previously, and are cited where appropriate.

Code availability

Computational methods used for this article—including functions for simulating birth–death models, for constructing models within a

given congruence class, for calculating the likelihood of a congruence class and for directly fitting congruence classes (either in terms of λ_p or in terms of r_p and $\rho\lambda_o$) to extant timetrees—are implemented in the R package *castor* v.1.5.5, which is available from The Comprehensive R Archive Network at <https://cran.r-project.org/package=castor>.

Acknowledgements S.L. was supported by a start-up grant by the University of Oregon. M.W.P. was supported by an NSERC Discovery Grant. We thank L. Harmon, S. Otto, A. MacPherson, D. Schluter, T. J. Davies, M. Whitlock, L. F. Henao Diaz, K. Kaur, J. Uyeda, D. Caetano, J. Rolland, L. Parfrey and A. Mooers for insightful comments on this work.

Author contributions S.L. performed the mathematical calculations and computational analyses. S.L. and M.W.P. conceived the project and contributed to the writing of the manuscript.

Competing interests The authors declare no competing interests.

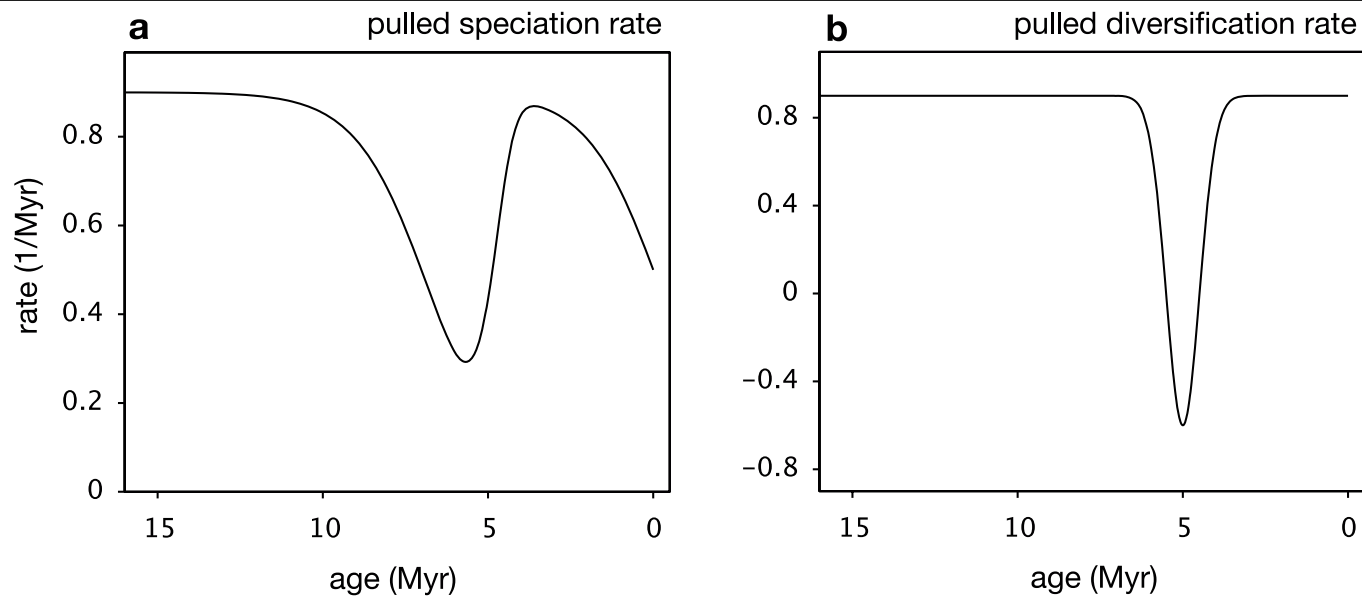
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2176-1>.

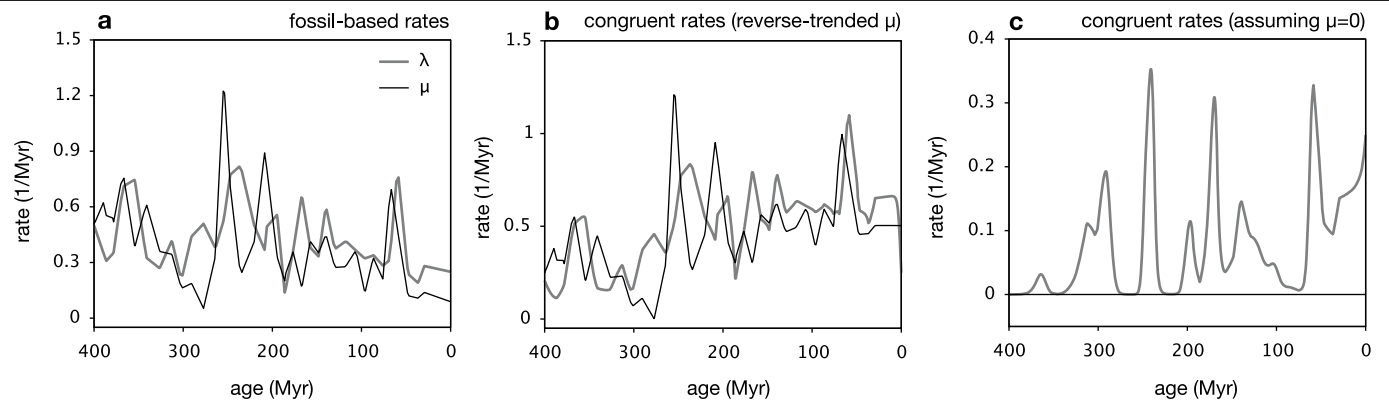
Correspondence and requests for materials should be addressed to S.L. or M.W.P.

Peer review information *Nature* thanks Lee Hsiang Liow, Antonis Rokas, Mike Steel and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

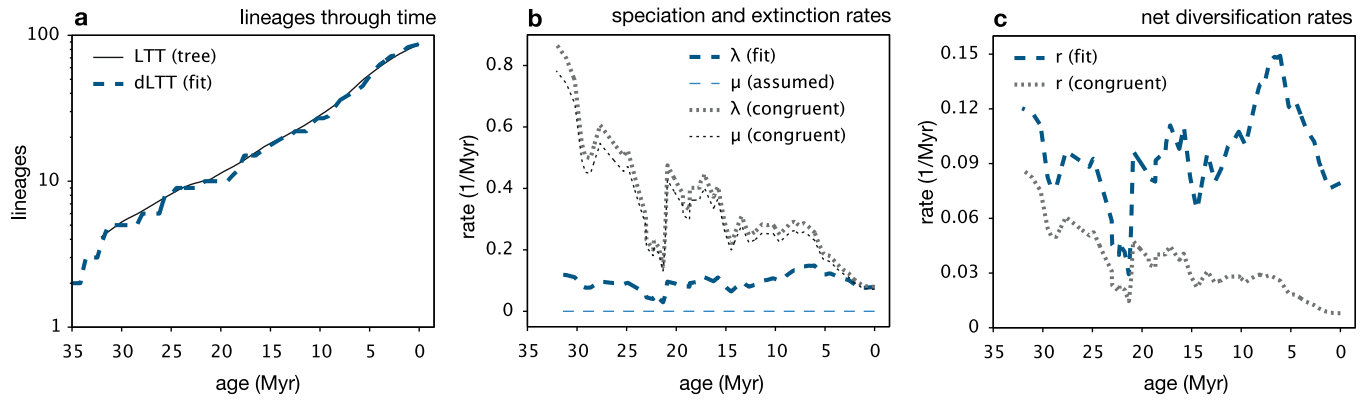


Extended Data Fig. 1 | Pulled speciation and diversification rates. a, b, Pulled speciation rate (a) and pulled diversification rate (b) of the four congruent models shown in Fig. 1.



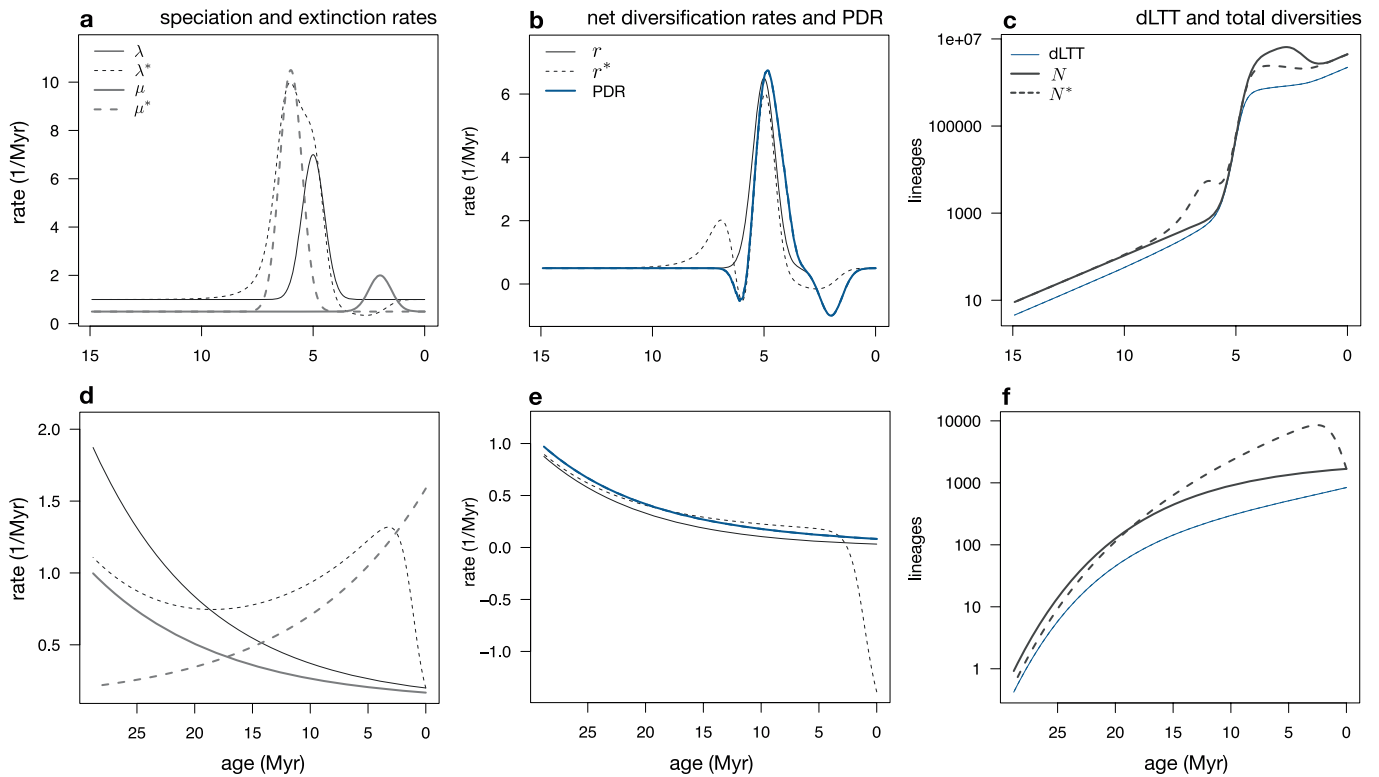
Extended Data Fig. 2 | Illustration of congruent birth–death processes (fossil data). **a**, Origination and extinction rates of marine invertebrate genera, estimated from fossil data. **b**, Congruent scenario to that in **a**, obtained by reversing the linear trend of μ (that is, fitting a linear curve to the original μ , and

then subtracting that curve twice) and adjusting λ according to equation (2). **c**, Congruent scenario to that in **a**, assuming an extinction rate of zero. Further details are provided in Supplementary Information section S.10.



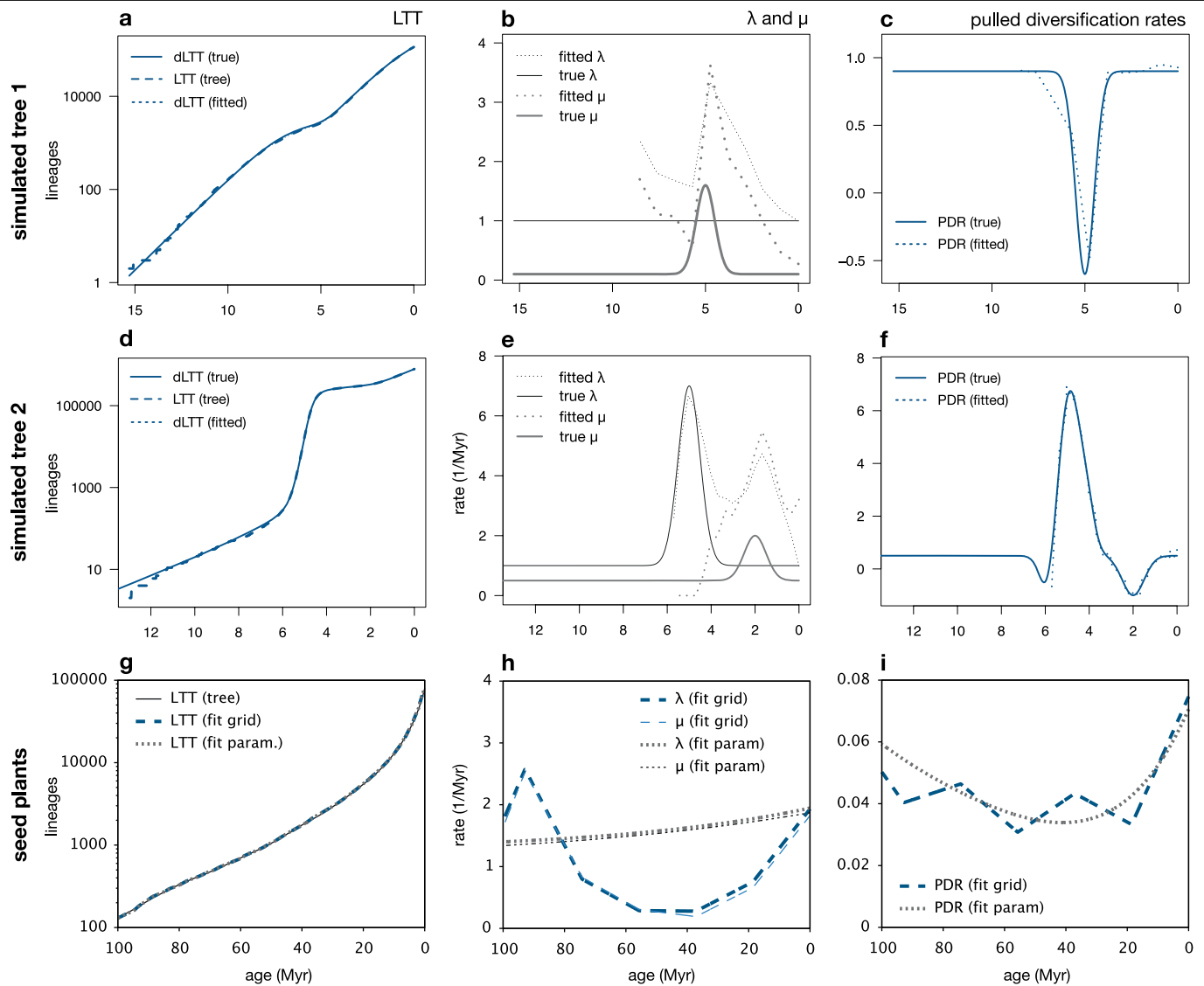
Extended Data Fig. 3 | Previous studies are likely to have over-interpreted phylogenetic data. Time-dependent birth–death model fitted to a nearly complete extant timetree of the Cetacea, under the assumption of extinction rates of zero ($\mu = 0$), compared to a congruent model in which the

rate is close to the speciation rate ($\mu = 0.9\lambda$). **a**, LTT of the tree, compared to the dLTT predicted by the two models. **b**, Speciation rates (λ) and extinction rates (μ) of the two models. **c**, Net diversification rates ($r = \lambda - \mu$) of the two models. Further details are provided in Supplementary Information section S.4.



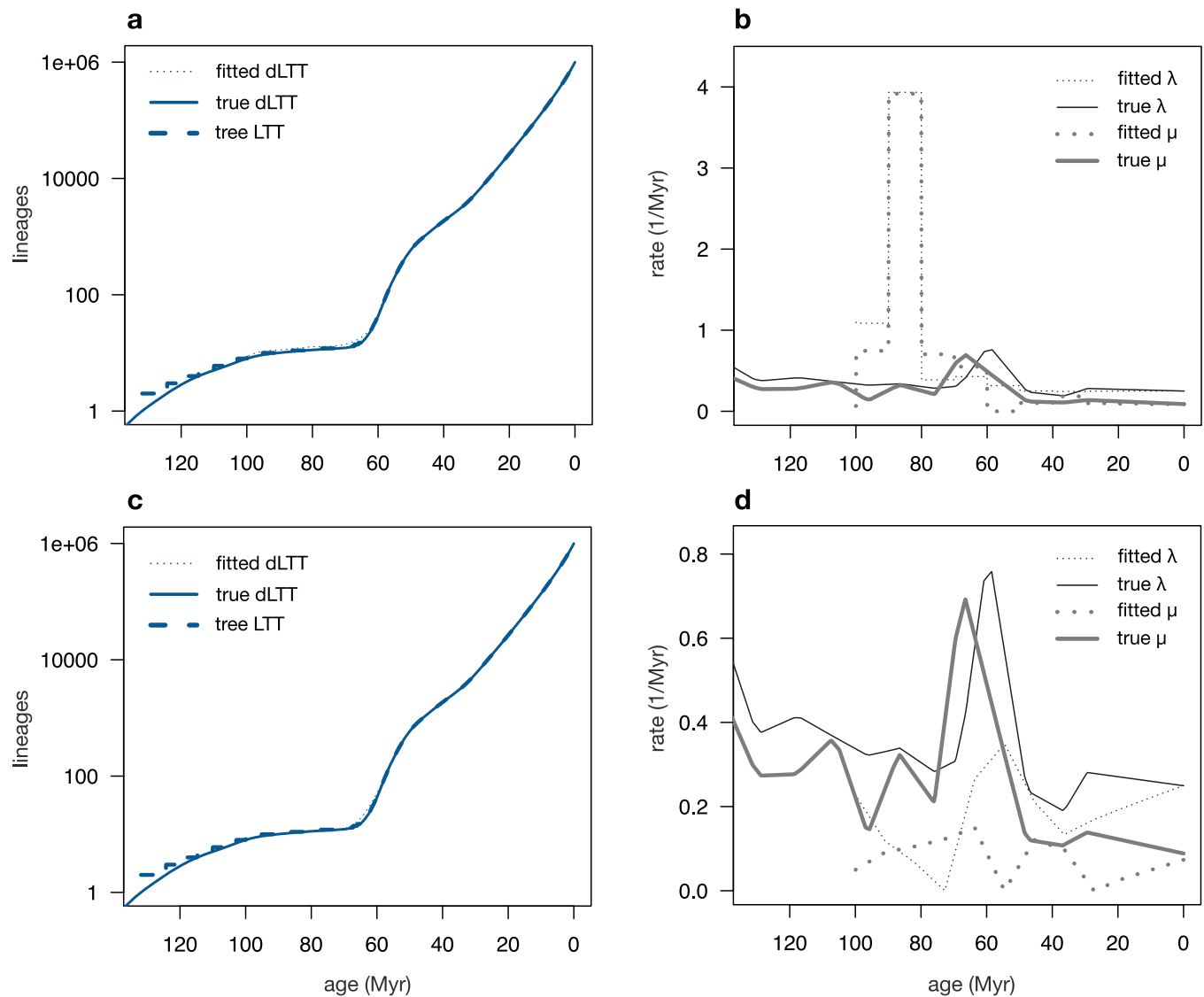
Extended Data Fig. 4 | Additional examples of congruent birth–death processes. **a–c**, Example of two congruent–yet markedly different–birth–death models. Both models exhibit a temporary spike in the extinction rate and a temporary spike in the speciation rate; however, the timings of these events differ substantially between the two models. Both models exhibit the same dLTT and the same pulled diversification rate (r_p) and would yield identical likelihoods for any given extant timetree. **a**, Speciation rates (λ and λ^*) and extinction rates (μ and μ^*) of the two models, plotted over time. Continuous

curves correspond to the first model, and dashed curves correspond to the second model. **b**, Net diversification rates (r and r^*) and pulled diversification rate (r_p) of the two models. **c**, dLTT and deterministic total diversities (N and N^*) predicted by the two models. **d–f**, Another example of two congruent models. In the first model, the speciation and extinction rates both decrease exponentially over time, whereas in the second model the extinction rate increases exponentially over time and the speciation rate exhibits variable directions of change over time. In all models, the sampling fraction is $\rho = 0.5$.



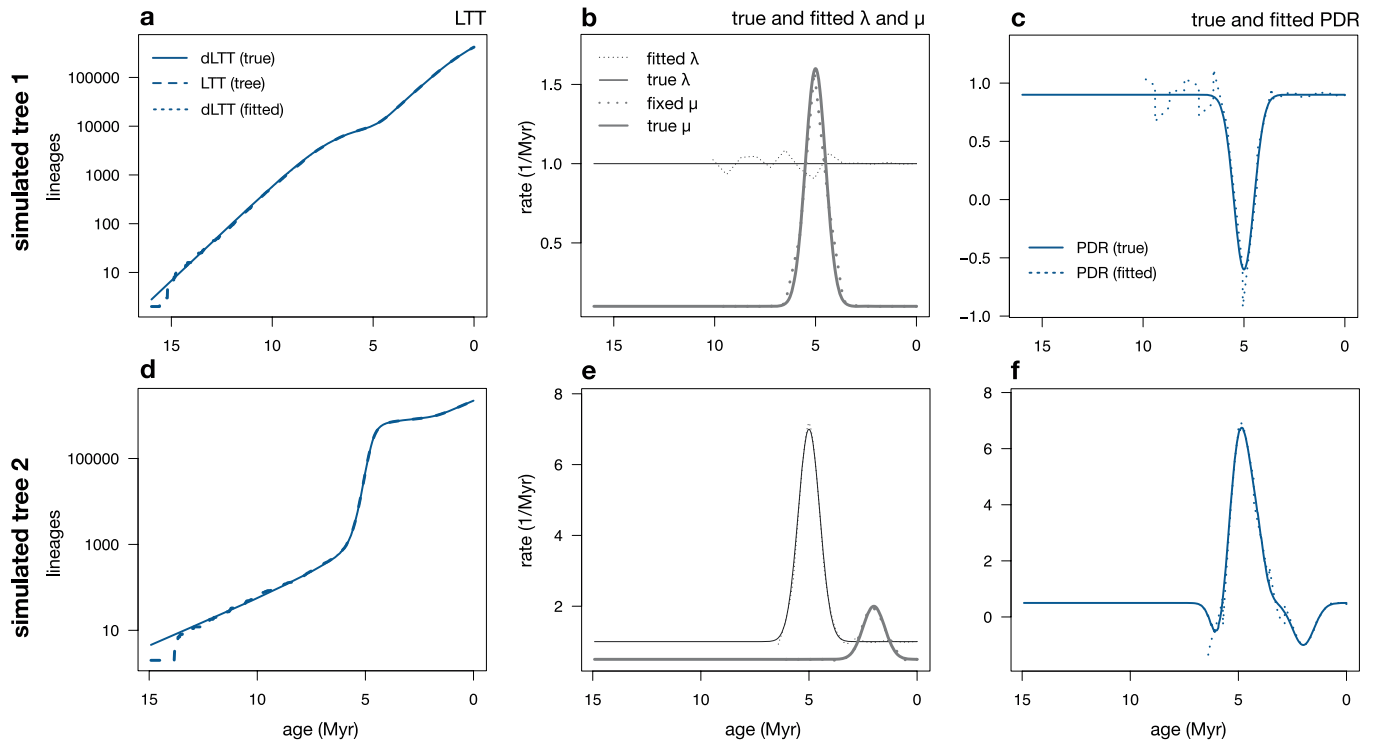
Extended Data Fig. 5 | Identifiability issues persist in large trees. **a–c**, Diversification analysis of a timetree (about 114,000 tips) simulated from a birth–death process that exhibits a mass extinction event at around 5 Myr before present. **a**, LTT of the generated tree (long-dashed curve), dLTT of the true model that generated the tree (continuous curve) and dLTT of a maximum-likelihood fitted model (short-dashed curve) are shown. The fitted dLTT is practically identical to the true dLTT and thus is covered by the latter. **b**, True speciation and extinction rates (continuous curves), compared to fitted speciation and extinction rates (dashed curves). There is considerable disagreement between the fitted and true λ and μ , despite the fact that the allowed model set could—in principle—approximate the true rates reasonably well. **c**, Pulled diversification rate (PDR) of the true model (continuous curve), compared to the pulled diversification rate of the fitted model (dashed curve). **d–f**, Diversification analysis of a timetree (about 785,000 tips) simulated from

a birth–death process that exhibits a rapid radiation event at around 5 Myr before present and a mass extinction event at around 2 Myr before present. **d–f** are analogous to **a–c**. There is considerable disagreement between the fitted and true λ and μ , despite the fact that the allowed model set could—in principle—approximate the true rates reasonably well. Extended Data Figure 7 provides the fitting results when μ is fixed to its true value. **g–i**, Diversification analyses of an extant timetree of 79,874 seed plant species, performed either by fitting λ and μ on a grid of discrete time points or by fitting the parameters of generic polynomial or exponential functions for λ and μ . **g**, LTT of the tree, dLTT of the grid-fitted model and dLTT of the fitted parametric model. **h**, Speciation and extinction rates predicted by the grid-fitted model or the fitted parametric model. **i**, Pulled diversification rate predicted by the grid-fitted model and the fitted parametric model. Further details are provided in Supplementary Information sections S.10 and S.11.



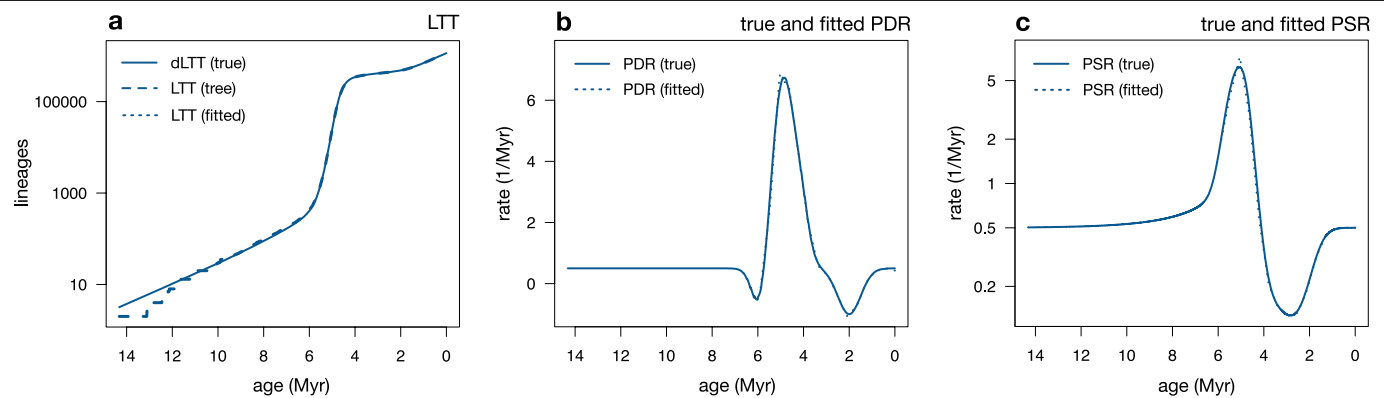
Extended Data Fig. 6 | Identifiability issues cannot be resolved with the Akaike information criterion. Maximum-likelihood birth–death models fitted to a tree comprising 1,000,000 tips, simulated on the basis of the origination and extinction rates of marine invertebrate genera estimated from fossil data. Top row, maximum-likelihood-fitted piecewise constant model (also known as birth–death–shift model), with grid size ($n=11$) chosen by minimizing the Akaike information criterion (AIC). Bottom row, maximum-likelihood-fitted piecewise linear model, with grid size ($n=12$) chosen by

minimizing the AIC. Left column, dLTTs of the fitted models compared to the true dLTT and the LTT of the tree. Right column, fitted speciation and extinction rates, compared to the true rates used to generate the tree. In both cases, the maximum-likelihood models poorly reflect the true rates despite a near-perfect match of the LTT, even when the complexity of the models was optimized on the basis of the AIC. For further details, see Supplementary Information sections S.2 and S.10.



Extended Data Fig. 7 | Estimating λ when μ and ρ are fixed or known. a–c, Example analysis of a simulated extant timetree (about 114,000 tips) that exhibits a mass extinction event at around 5 Myr before present. A birth–death model was fitted while fixing μ and ρ to their true values; λ was fitted at 15 discrete time points. **a**, LTT of the generated tree (long-dashed curve), dLTT of the true model that generated the tree (continuous curve) and dLTT of a maximum-likelihood fitted model (short-dashed curve). The fitted dLTT is practically identical to the true dLTT, and is thus covered by the latter. **b**, True speciation and extinction rates (continuous curves), along with the fitted speciation rate and fixed extinction rate (dashed curves). **c**, Pulled

diversification rate of the true model (r_p , continuous curve), compared to the pulled diversification rate of the fitted model (dashed curve). **d–f**, Example analysis of a simulated extant timetree (about 785,000 tips) that exhibits a rapid radiation event at about 5 Myr before present and a mass extinction event at about 2 Myr before present. A birth–death model was fitted similarly to the example shown in **a–c**, and **d–f** are analogous to **a–c**. In both cases, rate estimation was restricted to ages at which the LTT included at least 500 lineages. Further details are provided in Supplementary Information section S.10.



Extended Data Fig. 8 | Fitting congruence classes instead of models.

Analysis of an extant timetree generated by a birth–death model that exhibits a temporary rapid radiation event about 5 Myr before present and a mass extinction event about 2 Myr before present. A congruence class was fitted to the timetree either in terms of the pulled diversification rate (r_p) and the product $\rho\lambda_o$, or in terms of the pulled speciation rate (PSR) (λ_p). **a**, LTT of the tree (long-dashed curve), together with the dLTT of the true model (continuous curve) and the dLTT of the fitted congruence classes (short-dashed curve); in both cases, the fitted dLTT was almost identical to the true dLTT, and is thus

completely covered by the latter. **b**, Pulled diversification rate of the true model (continuous curve), compared to the fitted pulled diversification rate (short-dashed curve). **c**, Pulled speciation rate of the true model (continuous curve), compared to the fitted pulled speciation rate (short-dashed curve). The pulled diversification rate and pulled speciation rate were fitted via maximum-likelihood methods, allowing the pulled diversification rate or pulled speciation rate to vary freely at 15 discrete equidistant time points. Further details are provided in in Supplementary Information section S.9.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection No software was used to collect data. Simulated data was generated using the R package "castor", which is freely available on CRAN.

Data analysis Computational methods used for this article, including functions for simulating birth-death models, for constructing models within a given congruence class, for calculating the likelihood of a congruence class, and for directly fitting congruence classes to extant timetrees, are implemented in the R package "castor", which is freely available on CRAN.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

No raw data was generated for this manuscript. All phylogenetic datasets used have been published previously and are cited in the manuscript where appropriate.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	We use mathematical proofs, numerical simulations and analysis of previously published phylogenetic data to investigate the identifiability of past diversification dynamics from extant time-calibrated phylogenies.
Research sample	The following previously published datasets were used: Origination and extinction rates of marine invertebrate genera, estimated from the fossil record by (Alroy et al. 2008). Time-calibrated phylogeny of 79874 extant seed plant species, published by Smith and Brown (2018). Time-calibrated phylogeny of the Cetacea, published by Steeman et al. (2009).
Sampling strategy	Sample sizes (i.e., phylogeny sizes) were chosen as large as possible in our examples, to demonstrate that the identifiability issues discussed in the paper persist even for massive datasets. For simulated trees we used very large numbers of tips (larger than commonly seen in the literature), again to illustrate that our conclusions remain valid even for massive data sets.
Data collection	No new data were collected.
Timing and spatial scale	No new data were collected.
Data exclusions	No data was excluded from the analysis.
Reproducibility	Our full mathematical proofs and numerical procedures are described in detail in the manuscript and supplemental material. All new simulation code is published through the R package "castor" (version 1.5.5), which is freely available on CRAN.
Randomization	This is not relevant to our study, as no experiments were performed.
Blinding	Blinding was not relevant to our study, as no experiments were performed.
Did the study involve field work?	<input type="checkbox"/> Yes <input checked="" type="checkbox"/> No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Accurate compound-specific ^{14}C dating of archaeological pottery vessels

<https://doi.org/10.1038/s41586-020-2178-z>

Received: 24 May 2019

Accepted: 13 February 2020

Published online: 8 April 2020

 Check for updates

Emmanuelle Casanova¹, Timothy D. J. Knowles^{1,2}, Alex Bayliss^{3,4}, Julie Dunne¹, Marek Z. Barański⁵, Anthony Denaire⁶, Philippe Lefranc⁷, Savino di Lernia^{8,9}, Mélanie Roffet-Salque¹, Jessica Smyth^{1,10}, Alistair Barclay¹¹, Toby Gillard¹, Erich Claßen¹², Bryony Coles¹³, Michael Ilett¹⁴, Christian Jeunesse¹⁵, Marta Krueger¹⁶, Arkadiusz Marciniak¹⁶, Steve Minnitt¹⁷, Rocco Rotunno⁸, Pieter van de Velde¹⁸, Ivo van Wijk¹⁹, Jonathan Cotton²⁰, Andy Daykin²⁰ & Richard P. Evershed^{1,2}✉

Pottery is one of the most commonly recovered artefacts from archaeological sites. Despite more than a century of relative dating based on typology and seriation¹, accurate dating of pottery using the radiocarbon dating method has proven extremely challenging owing to the limited survival of organic temper and unreliability of visible residues^{2–4}. Here we report a method to directly date archaeological pottery based on accelerator mass spectrometry analysis of ^{14}C in absorbed food residues using palmitic ($\text{C}_{16:0}$) and stearic ($\text{C}_{18:0}$) fatty acids purified by preparative gas chromatography^{5–8}. We present accurate compound-specific radiocarbon determinations of lipids extracted from pottery vessels, which were rigorously evaluated by comparison with dendrochronological dates^{9,10} and inclusion in site and regional chronologies that contained previously determined radiocarbon dates on other materials^{11–15}. Notably, the compound-specific dates from each of the $\text{C}_{16:0}$ and $\text{C}_{18:0}$ fatty acids in pottery vessels provide an internal quality control of the results⁶ and are entirely compatible with dates for other commonly dated materials. Accurate radiocarbon dating of pottery vessels can reveal: (1) the period of use of pottery; (2) the antiquity of organic residues, including when specific foodstuffs were exploited; (3) the chronology of sites in the absence of traditionally datable materials; and (4) direct verification of pottery typochronologies. Here we used the method to date the exploitation of dairy and carcass products in Neolithic vessels from Britain, Anatolia, central and western Europe, and Saharan Africa.

Chronology lies at the heart of archaeology¹⁶. Radiocarbon dating by accelerator mass spectrometry (AMS) is the most widely used method for providing calendrical chronologies for human activities over the past 50,000 years¹⁷, and is most commonly performed on samples of charred plant remains and bone¹⁷. Radiocarbon dates can be used alongside relative sequences, such as those derived from stratigraphy or the typological analysis or seriation of artefact types, to build chronological models. Applying Bayes' theorem enables radiocarbon dating to provide calendar age estimates with uncertainties as low as a few decades¹⁸.

The invention of pottery in the late Pleistocene epoch was probably a critical driver for developments in food processing^{19,20}. Pottery vessels can often be placed in robust relative chronological sequences using typology and seriation, although obtaining precise and accurate

radiocarbon dates from pottery is challenging^{2,3,21}. All sources of carbon associated with pottery vessels have been considered for dating^{2–4}, including organic temper, which occasionally survives firing, and surficial food crusts, although these are rare and prone to contamination owing to their exposed nature²². By contrast, the lipidic components of food residues absorbed into—and protected by—the clay matrix during cooking occur very commonly⁸, often in high concentrations (milligrams per gram of clay fabric). These offer an untapped resource for radiocarbon dating. The most common absorbed residues correspond to degraded animal fats characterized by their high abundances of $\text{C}_{16:0}$ and $\text{C}_{18:0}$ fatty acids^{7,8}. The possibility of using preparative capillary gas chromatography (pcGC) to isolate chemically pure fatty acids from such residues for compound-specific radiocarbon analysis (CSRA) was recognized more than 20 years ago^{21,23,24}. Although initial attempts

¹Organic Geochemistry Unit, School of Chemistry, University of Bristol, Bristol, UK. ²Bristol Radiocarbon Accelerator Mass Spectrometry Facility, University of Bristol, Bristol, UK. ³Scientific Dating, Historic England, London, UK. ⁴Biological & Environmental Sciences, University of Stirling, Stirling, UK. ⁵Faculty of Architecture and Design, Academy of Fine Arts in Gdańsk, Gdańsk, Poland. ⁶University of Burgundy/UMR 6298 ARTEHIS, Dijon, France. ⁷University of Strasbourg UMR 7044/INRAP, Strasbourg, France. ⁸Dipartimento di Scienze dell'Antichità, Sapienza, Università di Roma, Rome, Italy. ⁹GAES, University of the Witwatersrand, Johannesburg, South Africa. ¹⁰School of Archaeology, University College Dublin, Dublin, Ireland. ¹¹Cotswold Archaeology, Cirencester, UK. ¹²LVR-State Service for Archaeological Heritage, Bonn, Germany. ¹³Department of Archaeology, University of Exeter, Exeter, UK. ¹⁴Université Paris 1 Panthéon-Sorbonne, UMR 8215 Trajectoires, Nanterre, France. ¹⁵University of Strasbourg, UMR7044, MISHA, Strasbourg, France. ¹⁶Institute of Archaeology, Adam Mickiewicz University, Poznań, Poland. ¹⁷Somerset County Museum, Taunton Castle, Taunton, UK. ¹⁸Archaeological Research Leiden, Leiden, The Netherlands. ¹⁹Faculty of Archaeology, Leiden University, Leiden, The Netherlands. ²⁰Museum of London Archaeology (MOLA), London, UK. ✉e-mail: r.p.evershed@bristol.ac.uk

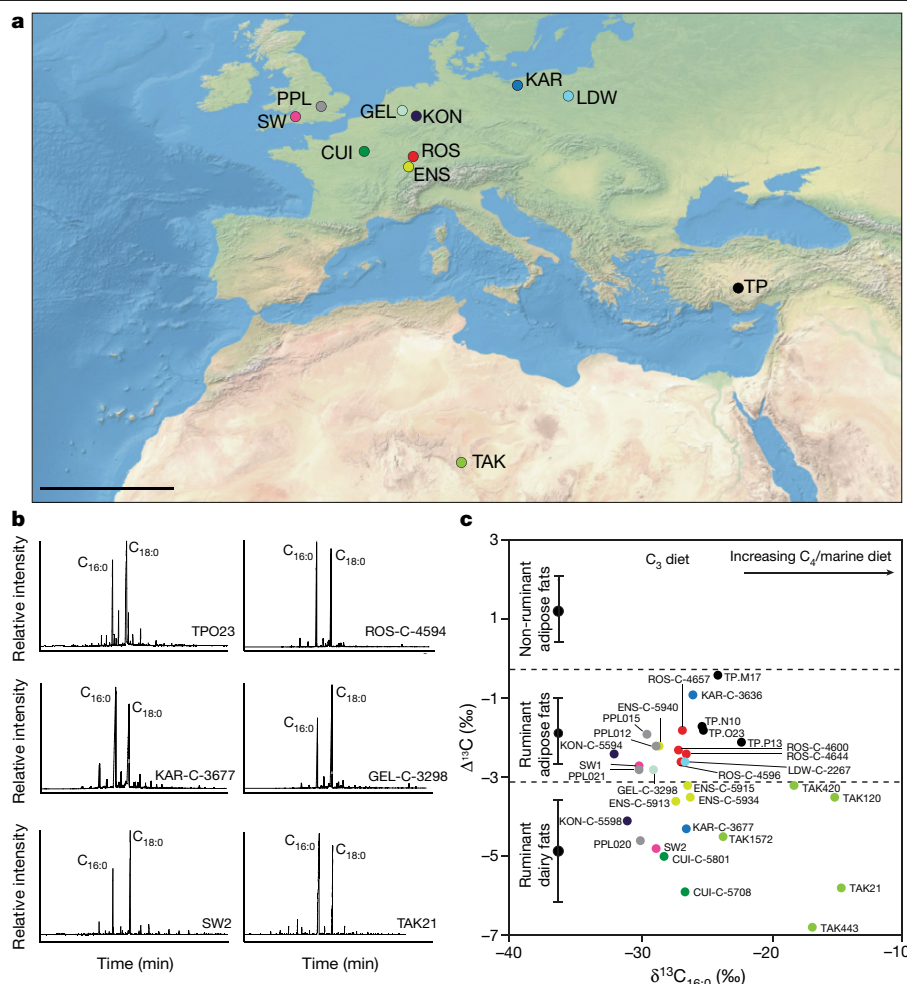


Fig. 1 | Site location map, partial gas chromatograms and stable isotope determination of compound-specific radiocarbon-dated lipid residues preserved in Neolithic pottery vessels. a, Map of the location of the archaeological sites for which CSRA was used in this study. Scale bar, 1,000 km. CUI, Cuiry-lès-Chaudardes; ENS, Ensisheim; GEL, Geleen-Janskamperveld; KAR, Karwowo 1; KON, Königshoven 14; LDW, Ludwinowo 7; PPL, Principal

Place, London; ROS, Rosheim; SW, Sweet Track; TAK, Takarkori; TP, Çatalhöyük East. **b**, Partial gas chromatograms of a selection of potsherds showing C_{16:0} and C_{18:0} fatty acid abundances. **c**, Scatter plots of $\Delta^{13}\text{C}$ ($=\delta^{13}\text{C}_{18:0}-\delta^{13}\text{C}_{16:0}$) values plotted against $\delta^{13}\text{C}_{16:0}$ values (mean of 2 measurements) for all of the sherds dated ($n=31$), ranges on the left denote the mean \pm 1 s.d. of modern reference fats, as reported in ref. ²⁸.

to date pottery vessels were promising, the accuracy and precision demanded by archaeology could not be achieved owing to unidentified technical difficulties, leading to highly variable results^{21,23}.

We have brought together the latest technologies for radiocarbon measurements, including automated graphitization and MICADAS compact AMS, in conjunction with high-field 700-MHz NMR, to undertake systematic investigations of the pcGC protocol^{5,6}. Rigorous assessment of contamination in compounds purified by pcGC was undertaken, leading to our invention of a solventless pcGC trap and implementation of cleaning procedures to avoid between-run carryover^{5,6}. These advances reduce the exogenous contamination of fatty acids that has previously been associated with pcGC to below concentrations that would significantly affect measured radiocarbon ages. For archaeological animal fats, it has previously been demonstrated that two fatty acids isolated from the same matrix generate the same radiocarbon age (that is, statistically consistent at the 95% significance level), providing an internal quality control for archaeological dating⁶. In this study, we aim to extend this method to archaeological pot lipids. We selected pottery vessels that were rich in animal fats from our database of lipid residues that we accumulated over the last three decades. Pottery vessels from chronologically well-characterized settings and different burial environments were analysed and the compatibility of pot lipid dates with

these existing chronologies was evaluated by statistical comparison of posterior density estimates for the key parameters and the use of indices of agreement with inclusion in these known frameworks (Fig. 1, Extended Data Table 1 and Supplementary Information 1).

We initially focused on Neolithic Carinated Bowl pottery from the Sweet Track (Fig. 2a), an elevated wooden trackway discovered in a wetland area of the Somerset Levels^{9,10,25} in the United Kingdom (Supplementary Information 2). This site is critical because its construction has been precisely dated by dendrochronology to the winter–early spring of 3807–3806 BC and the trackway was used and maintained for approximately 10 years¹⁰. Lipids from pots that were found alongside the trackway, and were probably contemporaneous to its construction and use, have previously been dated, but the measured dates were a century later than the construction of the trackway²³. Re-analysis of the two vessels (Fig. 2b) using our new approach produced uncalibrated radiocarbon ages of $5,110 \pm 25$ years before present (BP; taken as AD 1950) (SW1) and $5,092 \pm 26$ BP (SW2), which are statistically indistinguishable ($T=9.0$, $T'(5\%)=9.5$, $v=4$) from the measurements of the tree rings included in the IntCal13 calibration curve for the relevant decade²⁶ (Fig. 2c). The calibrated dates of these ages are clearly compatible with the tree-ring dates for the construction of the trackway.

Extending our approach to Anatolia, the Neolithic tell of Çatalhöyük East was a locus for the emergence and development of pottery

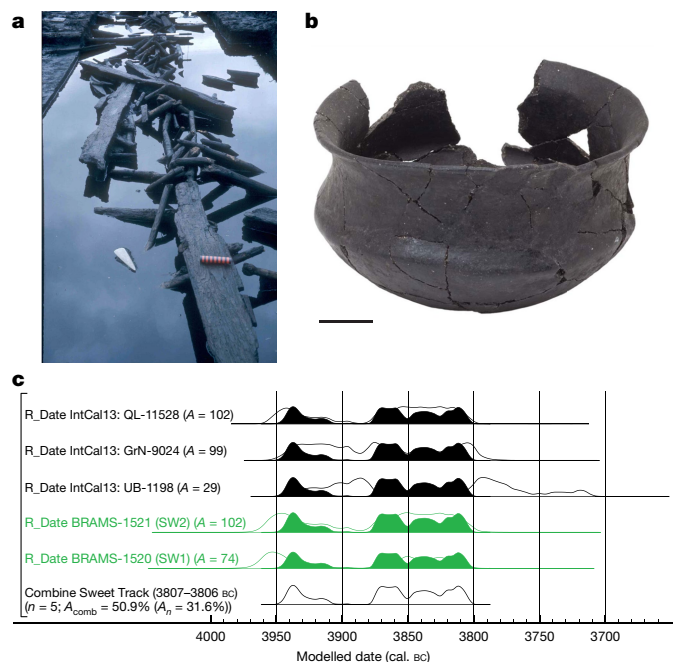


Fig. 2 | Sweet Track timbers, a pottery vessel and calibrated radiocarbon dates. **a**, Photograph of Sweet Track timbers. **b**, Photograph of a Carinated Bowl (SW2) that was recovered alongside the Sweet Track. Scale bar, 5 cm. **c**, Probability distributions of dates from pots deposited next to the Sweet Track (green) and from oak trees (black) included in IntCal13²⁶ that include the date of the Sweet Track construction in 3807–3806 BC. Each distribution represents the relative probability that an event occurs at a particular time. For each of the dates, two distributions have been plotted: one in outline, which is the simple radiocarbon calibration, and a solid distribution, based on the model used. The square bracket down the left side along with the OxCal keywords define the overall model exactly (provided in Supplementary Information 2). A , A_{comb} and A_n are the individual agreement indices, the combination agreement indices and the acceptable threshold to combine n radiocarbon dates, respectively. The photographs were provided by S.M. and are reproduced with permission from the Somerset Levels Project.

production. A 21-m-deep stratigraphic sequence provides strong archaeological prior information for a Bayesian chronological model that covers the upper parts of the mound (TP area)¹¹. The sequence of houses, middens and burial structures has been combined with 50 radiocarbon dates, revealing a Neolithic sequence of occupation from the mid-sixty-fourth to the mid-sixtieth centuries calibrated (cal.) BC¹¹. Our compound-specific radiocarbon ages on adipose lipids²⁷ from four pottery vessels from four different contexts (TP.M17, $7,382 \pm 31$ BP; TP.N10, $7,348 \pm 25$ BP; TP.O23, $7,340 \pm 27$ BP; and TP.P13, $7,364 \pm 25$ BP) were incorporated into the Bayesian chronological model for this part of the site (Extended Data Figs. 1, 2 and Supplementary Information 3). The revised model for the Neolithic deposits in the TP area shows posterior distributions for the key parameters that are almost identical to those from the original model¹¹. Their median values vary by an average of 4 years and a maximum of 10 years, confirming the compatibility of the radiocarbon ages determined using fatty acids with the site stratigraphy. On the basis of sensitivity analyses (Supplementary Information 3), this well-constrained model is at least as sensitive as measurements on paired materials to detect inaccuracies. In this case, the CSRA dates not only provide direct dating for the importance of ruminant carcass products (Fig. 1c) to the inhabitants of Çatalhöyük at this time (derived from $\delta^{13}\text{C}$ values of preserved fats), but also provide direct dating evidence for the climatic changes associated with the global event of 8.2 thousand years ago (derived from compound-specific deuterium isotope analyses using the same fats)²⁷.

The next analysis tests the accuracy of our dating approach using a classic pottery seriation study related to Neolithic ceramics from Lower Alsace (France) that spans the second quarter of the fifth millennium cal. BC¹² (Supplementary Information 4). The regional correspondence analysis clearly separates the Hinkelstein, Grossgartach, Planig-Friedberg and Rössen Middle Neolithic ceramic groups. We focused on vessels from three pits, all of which can be assigned to the Grossgartach phase (Fig. 3a, b). The sequence of ceramic phases was combined with the existing assemblage of 95 radiocarbon dates, which were largely measured on articulated bones, along with four CSRA dates on fatty acids (ROS-C-4596, $5,804 \pm 25$ BP; ROS-C-4600, $5,904 \pm 28$ BP; ROS-C-4644, $5,931 \pm 26$ BP; and ROS-C-4657, $5,912 \pm 28$ BP) from the Grossgartach sherds in a model using Bayesian statistics. The phase boundaries in this revised model are very similar to those produced by the original analysis¹², as median values differ by an average of 6 years and a maximum of 15 years (Fig. 3c). The sensitivity analyses (Supplementary Information 4) demonstrate that the model is particularly sensitive to small biases, and probably more sensitive than measurements on paired materials. The CSRA dates are clearly compatible with the attribution of these pottery vessels to the Grossgartach ceramic phase based on their decorative motifs, and with the other radiocarbon dates for this group.

We then explored the introduction of a new food product—that is, milk—into Neolithic Europe by undertaking radiocarbon dating of animal fat residues, including dairy fats, that were recovered from early farming settlements with Linearbandkeramik (LBK) pottery (Fig. 1). These communities settled in central Europe from the early fifty-fourth century BC¹³. Animal fats in 12 potsherds from the earliest LBK contexts at 6 sites, in Poland, France, Germany and the Netherlands, produced radiocarbon dates that were modelled and shown to be compatible with the currency of LBK ceramics in northern and western Europe^{12,13} (Extended Data Figs. 3, 4 and Supplementary Information 5). Sensitivity analyses (Extended Data Fig. 4 and Supplementary Information 5) demonstrate that this model is more sensitive to older biases as we focused on early settlements, illustrating the direct dating of a new food commodity. The radiocarbon dates on the earliest dairying residues suggest that the practice began in 5385–5225 cal. BC (95% probability; start LBK lipid; Extended Data Fig. 3) and probably arrived with the earliest farmers in these areas. Thus, the linking of fatty acid structures with compound-specific carbon isotope values and CSRA dates provides a powerful means of directly dating prehistoric foodways and their introduction.

We next investigated pottery from the Sahara Desert to provide a test of the methodology for a region in which depositional conditions are very different from the temperate climes of northern Europe. The Takarkori rock shelter, located in the now hyper-arid area of the Acacus Mountains, southwest Libya, demonstrates evidence of animal exploitation based on rock art and archaeological finds¹⁴ (Extended Data Figs. 5, 6). Previous work revealed abundant adipose and dairy fat residues in fragments of the pottery vessels²⁸. Stratigraphy and radiocarbon dating of a range of materials (bone collagen, charred plant remains, dung, skin and enamel bioapatite) placed deposits associated with Middle Pastoral pottery in the sixth–fifth millennia cal. BC^{14,28,29}. The fatty acids from 5 potsherds, containing dairy fat (Extended Data Fig. 6b), produced uncalibrated radiocarbon ages of $5,993 \pm 28$ BP (TAK443), $5,979 \pm 28$ BP (TAK120), $5,493 \pm 28$ BP (TAK420), $5,348 \pm 24$ BP (TAK21) and $5,085 \pm 24$ BP (TAK1572). The CSRA dates were proven to be entirely compatible with the currency of Middle Pastoral Neolithic ceramics (Extended Data Fig. 6d and Supplementary Information 6), and the direct radiocarbon dating of dairy residues confirms that dairying in North Africa began as early as the end of the sixth millennium cal. BC^{14,28,29}. Although the model sensitivity is weak based on the small number of reference dates that it includes (Extended Data Fig. 7 and Supplementary Information 6), it demonstrates the possibility of dating potsherds from extremely arid burial conditions. In addition, direct dating of pottery lipids represents

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2178-z>.

- Orton, C. & Hughes, M. *Pottery in Archaeology* 2nd edn (Cambridge Univ. Press, 2014).
- Evin, J., Gabasio, M. & Lefevre, J. C. Preparation techniques for radiocarbon dating of potsherds. *Radiocarbon* **31**, 276–283 (1989).
- Hedges, R. M., Tiemei, C. & Housley, R. A. Results and methods in the radiocarbon dating of pottery. *Radiocarbon* **34**, 906–915 (1992).
- Gabasio, M., Evin, J., Arnal, G. B. & Andrieux, P. Origins of carbon in potsherds. *Radiocarbon* **28**, 711–718 (1986).
- Casanova, E., Knowles, T. D. J., Williams, C., Crump, M. P. & Evershed, R. P. Use of a 700 MHz NMR microcryoprobe for the identification and quantification of exogenous carbon in compounds purified by preparative capillary gas chromatography for radiocarbon determinations. *Anal. Chem.* **89**, 7090–7098 (2017).
- Casanova, E., Knowles, T. D. J., Williams, C., Crump, M. P. & Evershed, R. P. Practical considerations in high-precision compound-specific radiocarbon analyses: eliminating the effects of solvent and sample cross-contamination on accuracy and precision. *Anal. Chem.* **90**, 11025–11032 (2018).
- Evershed, R. P. et al. Chemistry of archaeological animal fats. *Acc. Chem. Res.* **35**, 660–668 (2002).
- Roffet-Salque, M. et al. From the inside out: upscaling organic residue analyses of archaeological ceramics. *J. Archaeol. Sci. Rep.* **16**, 627–640 (2017).
- Coles, J. M. & Orme, B. J. Ten excavations along the Sweet Track (3200 bc). *Somerset Lev. Pap.* **10**, 5–45 (1984).
- Hillam, J. et al. Dendrochronology of the English Neolithic. *Antiquity* **64**, 210–220 (1990).
- Marciniak, A. et al. Fragmenting times: interpreting a Bayesian chronology for the late Neolithic occupation of Çatalhöyük East, Turkey. *Antiquity* **89**, 154–176 (2015).
- Denaire, A. et al. The cultural project: formal chronological modelling of the early and middle Neolithic sequence in Lower Alsace. *J. Archaeol. Method Theory* **24**, 1072–1149 (2017).
- Jakucs, J. et al. Between the Vinča and Linearbandkeramik worlds: the diversity of practices and identities in the 54th–53rd centuries cal bc in Southwest Hungary and beyond. *J. World Prehist.* **29**, 267–336 (2016).
- Biagetti, S. & di Lernia, S. Holocene deposits of Saharan rock shelters: the case of Takarkori and other sites from the Tadrart Acacus Mountains (southwest Libya). *Afr. Archaeol. Rev.* **30**, 305–338 (2013).
- Whittle, A. W. R., Healy, F. M. A. & Bayliss, A. *Gathering Time: Dating the Early Neolithic Enclosures of Southern Britain and Ireland* (Oxbow Books, 2011).
- Wheeler, R. E. M. *Archaeology from the Earth* (Penguin, 1956).
- Taylor, R. E. *Radiocarbon Dating, An Archaeological Perspective* (Academic, 1987).
- Bronk Ramsey, C. Bayesian analysis of radiocarbon dates. *Radiocarbon* **51**, 337–360 (2009).
- Barnett, W. & Hoopes, J. W. *The Emergence of Pottery: Technology and Innovation in Ancient Societies* (Smithsonian Institution Press, 1995).
- Kuzmin, Y. The origins of pottery in East Asia: updated analysis (the 2015 state-of-the-art). *Doc. Praehist.* **42**, 1–11 (2015).
- Stott, A. W. et al. Radiocarbon dating of single compounds isolated from pottery cooking vessel residues. *Radiocarbon* **43**, 191–197 (2001).
- Evershed, R. P. Biomolecular archaeology and lipids. *World Archaeol.* **25**, 74–93 (1993).
- Berstan, R. et al. Direct dating of pottery from its organic residues: new precision using compound-specific carbon isotopes. *Antiquity* **82**, 702–713 (2008).
- Eglinton, T. I., Aluwihare, L. I., Bauer, J. E., Druffel, E. R. M. & McNichol, A. P. Gas chromatographic isolation of individual compounds from complex matrices for radiocarbon dating. *Anal. Chem.* **68**, 904–912 (1996).
- Coles, B. J. & Coles, J. M. *Sweet Track to Glastonbury: The Somerset Levels in Prehistory* 163–169 (Oxbow, 1986).
- Reimer, P. J. et al. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal bp. *Radiocarbon* **55**, 1869–1887 (2013).
- Roffet-Salque, M. et al. Evidence for the impact of the 8.2-kyBP climate event on Near Eastern early farmers. *Proc. Natl Acad. Sci. USA* **115**, 8705–8709 (2018).
- Dunne, J. et al. First dairying in green Saharan Africa in the fifth millennium bc. *Nature* **486**, 390–394 (2012).
- Cherkinsky, A. & di Lernia, S. Bayesian approach to ¹⁴C dates for estimation of long-term archaeological sequences in arid environments: the Holocene site of Takarkori Rockshelter, Southwest Libya. *Radiocarbon* **55**, 771–782 (2013).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Lipid extraction and isolation

Potsherds were selected on the basis of the presence of terrestrial animal fats (dairy and ruminant carcass fats) in the lipid residue to avoid any possible reservoir effect caused by the processing of aquatic products in pots. A piece of 1–10 g of the potsherd was sampled, according to the lipid concentration. The sherds were extracted in a glass culture tube using H₂SO₄/MeOH (4% v/v, 3 × 8 ml, 70 °C, 1 h). The supernatants were centrifuged (2,500 rpm, 10 min) and combined into new culture tubes containing double-distilled water (5 ml). The lipids were extracted with *n*-hexane (4 × 5 ml), transferred into 3.5-ml vials and blown to dryness at room temperature under a gentle nitrogen stream. Subsequently, around 180 µl of *n*-hexane was added to obtain a concentration of fatty acid methyl esters (FAMES) at 5 µg of C µl⁻¹ before transfer to an autosampler vial for isolation by pcGC.

The pcGC consisted of a Hewlett Packard 5890 series II gas chromatograph coupled to a Gerstel Preparative Fraction Collector by a heated transfer line. The pcGC was equipped with a column with a 100% poly(dimethylsiloxane) stationary phase (Rxi-1ms, 30 m × 0.53 mm inner diameter, 1.5 µm film thickness, Restek). Helium was used as the carrier gas at a constant pressure of 10 psi. The GC temperature programme started with an isothermal hold at 50 °C for 2 min, the temperature was increased to 200 °C at 40 °C min⁻¹, to 270 °C at 10 °C min⁻¹ and finally increased to 300 °C at 20 °C min⁻¹ and held for 8.75 min. The C_{16:0} and C_{18:0} FAMES were injected (1 µl per run), separated and trapped 40 times per trapping sequence. Of the GC column effluent, 1% flows to the flame ionization detector, while the remaining 99% passes through a transfer line into the fraction collector, both of which were heated to 300 °C. Compounds were isolated based on their retention times⁶. The stationary phase degradation of the pcGC column and other sources of exogenous carbonaceous contamination were monitored on a Bruker Avance III HD 700 MHz NMR instrument following a previously published procedure^{5,6}.

Radiocarbon determinations and statistical analysis

The pcGC isolated compounds were transferred into Al capsules, after which they were combusted and graphitized in a Vario Microcube Elemental Analyser linked to an Automated Graphitisation System (AGE 3, IonPlus). All of the radiocarbon measurements were performed by the Bristol Accelerator Mass Spectrometer (BRAMS) facility at the University of Bristol. Data reduction was performed using the software BATS³⁰ (v.4.07). Radiocarbon dates obtained for FAMES were corrected for the presence of added methyl carbon using a mass balance approach^{5,6,21} and reported as the conventional radiocarbon ages³¹ (Supplementary Information 1).

Two contemporaneous compounds (C_{16:0} and C_{18:0} fatty acids) were dated and every pair of statistically indistinguishable measurements (at the 95% significance level)³² was combined as a weighted average before Bayesian chronological modelling using OxCal v.4.2 and v.4.3^{18,33} and the currently internationally agreed radiocarbon calibration curve for the Northern Hemisphere, IntCal13²⁶. The compatibility of the radiocarbon dates on absorbed fatty residues with existing sites and regional chronologies was assessed by including the lipid radiocarbon dates into existing statistical frameworks in a position defined by archaeological information (for example, stratigraphy or seriation). Their compatibility with the existing chronologies were achieved by: (1) comparison of posterior density estimates for key modelled parameters with equivalent date estimates or known age by dendrochronology; (2) using the individual and model agreement indices^{18,33} in models containing fatty acid dates; and (3) comparing posterior density estimates for key parameters from models that include the fatty acid dates to a model that does not include the fatty acid dates (Supplementary

Information 1). The sensitivity of existing chronological models to the addition of the new radiocarbon measurements was evaluated as above, after deliberately biasing the radiocarbon dates on pottery vessels to varying degrees while assessing the effect on posterior density estimates for the key parameters and indices of agreements (Supplementary Information 1–7).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data generated during this study are included in the Article, Extended Data Figs. 1–9, Extended Data Table 1 and Supplementary Information.

Code availability

The codes used in OxCal for statistical modelling are provided in the Supplementary Information.

30. Wacker, L., Christl, M. & Synal, H.-A. BATS: a new tool for AMS data reduction. *Nucl. Instrum. Methods Phys. Res. B* **268**, 976–979 (2010).
31. Stuiver, M. & Polach, H. A. Discussion reporting of ¹⁴C data. *Radiocarbon* **19**, 355–363 (1977).
32. Ward, G. K. & Wilson, S. R. Procedures for comparing and combining radiocarbon age determinations: a critique. *Archaeometry* **20**, 19–31 (1978).
33. Bronk Ramsey, C. Radiocarbon calibration and analysis of stratigraphy: the OxCal program. *Radiocarbon* **37**, 425–430 (1995).
34. Stuiver, M. & Reimer, P. J. Extended ¹⁴C data base and revised CALIB 3.0 ¹⁴C age calibration program. *Radiocarbon* **35**, 215–230 (1993).

Acknowledgements We thank the European Research Council for funding an advanced grant (NeoMilk, FP7-IDEAS-ERC/324202) and a proof-of-concept grant (LipDat, H2020 ERC-2018-PoC/812917) to R.P.E., financing a PhD to E. Casanova and postdoctoral contract to M.R.-S. and J.S., and a postdoctoral contract to E. Casanova; the BRAMS facility for the radiocarbon measurements, establishment of which was jointly funded by the NERC, BBSRC and University of Bristol; P. Monaghan for his help with the radiocarbon sample preparation; the Polish National Science Centre (decision DEC-2012/06/M/H3/00286) for financing the work in the upper levels at Çatalhöyük; the Department of Antiquities in Tripoli, Libya for permits and Sapienza University of Rome and Italian Ministry of Foreign Affairs for funding the fieldwork in Libya; MOLA (Museum of London Archaeology) for excavating and providing potsherds from Principal Place (PPL11), London EC2/E1; and B. Schnitzler from the Palais Rohan for accessing the material from Rosheim, A. Mulot from the Achéologie Alsace (Centre of Conservation and Study) for accessing the material from Ensisheim, R. W. Schmitz from the LVR-Landes Museum Bonn for accessing the material from Königshoven 14 and R. Brunning from the South West Heritage Trust for sharing excavation photographs of the Sweet Track.

Author contributions R.P.E. conceived the project. E. Casanova, R.P.E. and A. Bayliss wrote the paper. E. Casanova, T.D.J.K. and R.P.E. developed the method for dating lipids. E. Casanova, J.D. and T.G. performed the preparation of pottery vessels for radiocarbon analysis. E. Casanova and T.D.J.K. generated the radiocarbon measurements and performed data analysis. A. Bayliss undertook the statistical modelling of the radiocarbon dates. M.Z.B. advised on the stratigraphic sequence of the TP area of Çatalhöyük East. A.M. performed the stratigraphic analysis of the TP area of Çatalhöyük East and chronological analysis of the LBK from the Polish lowlands and M.K. helped with the selection and provided the pottery vessels from these sites. C.J. and P.L. excavated the Alsatian sites. A. Denaire and P.L. performed the correspondence analysis of the Alsace region. S.d.L. advised on the stratigraphic sequence and pottery analysis of Takarkori and R.R. studied the pottery assemblage. E. Casanova, M.R.-S. and J.S. sampled the LBK sites. M.R.-S. coordinated and processed the analyses of sherds from the LBK culture and from Çatalhöyük East. A. Barclay advised on project design. B.C. directed excavations of the Sweet Track and S.M. provided the pottery vessels. E. Claßen analysed the material and advised sampling for Königshoven 14. M.I. excavated and provided vessels from Cuiry-lès-Chaudardes. I.v.W. excavated and advised sampling from The Netherlands and P.v.d.V. analysed the material. A. Daykin excavated the site of Principal Place, London EC2/E1 as project manager and J.C. studied the pottery material.

Competing interests The authors declare no competing interests.

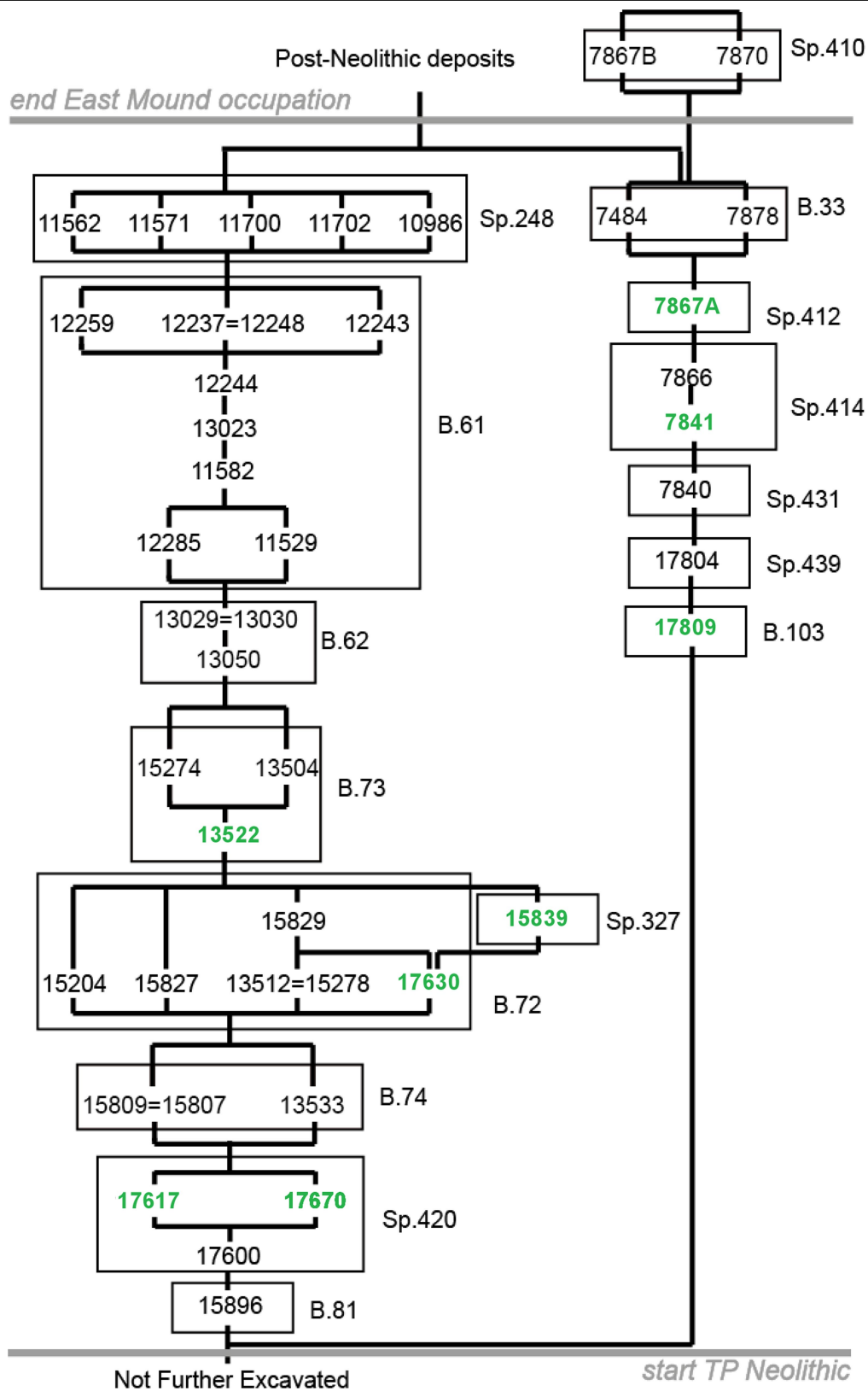
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2178-z>.

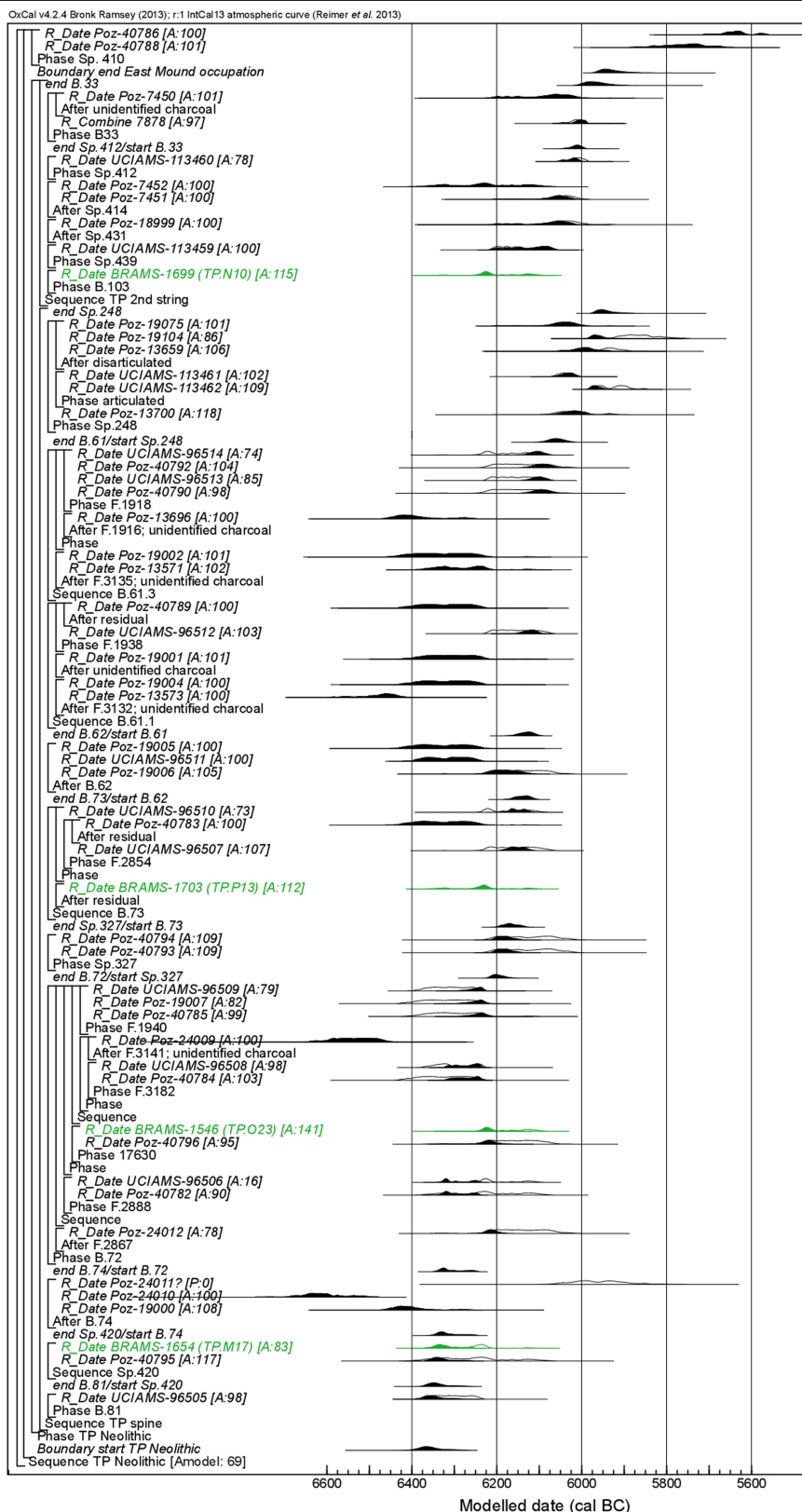
Correspondence and requests for materials should be addressed to R.P.E.

Peer review information Nature thanks Graeme Barker and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

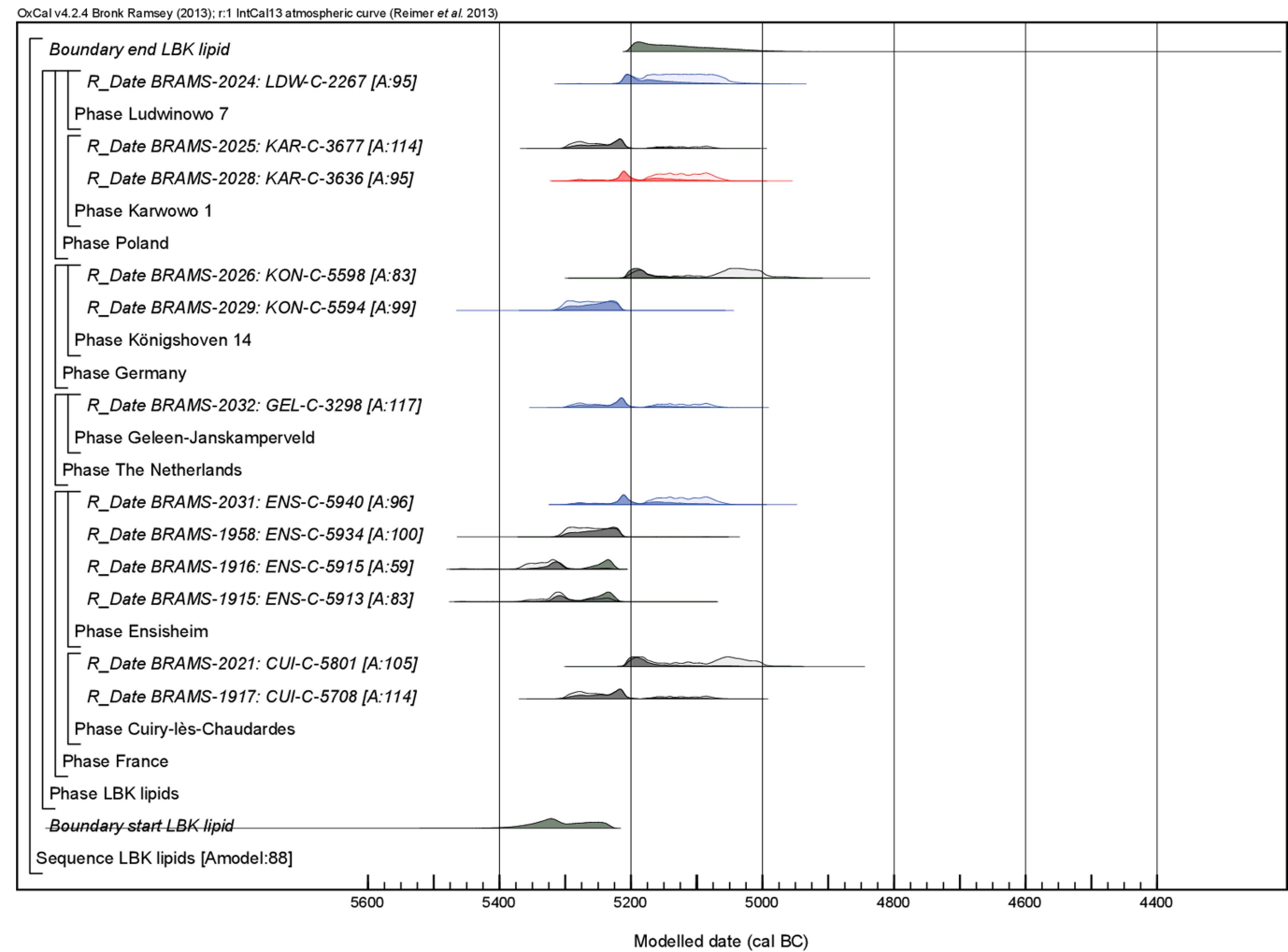


Extended Data Fig. 1 | Schematic showing the stratigraphic information of the Neolithic occupation of the TP area at Catalhöyük (Turkey). This information was included in the chronological model defined in Extended Data Fig. 2. Contexts containing potsherds dated in this study are highlighted in green.

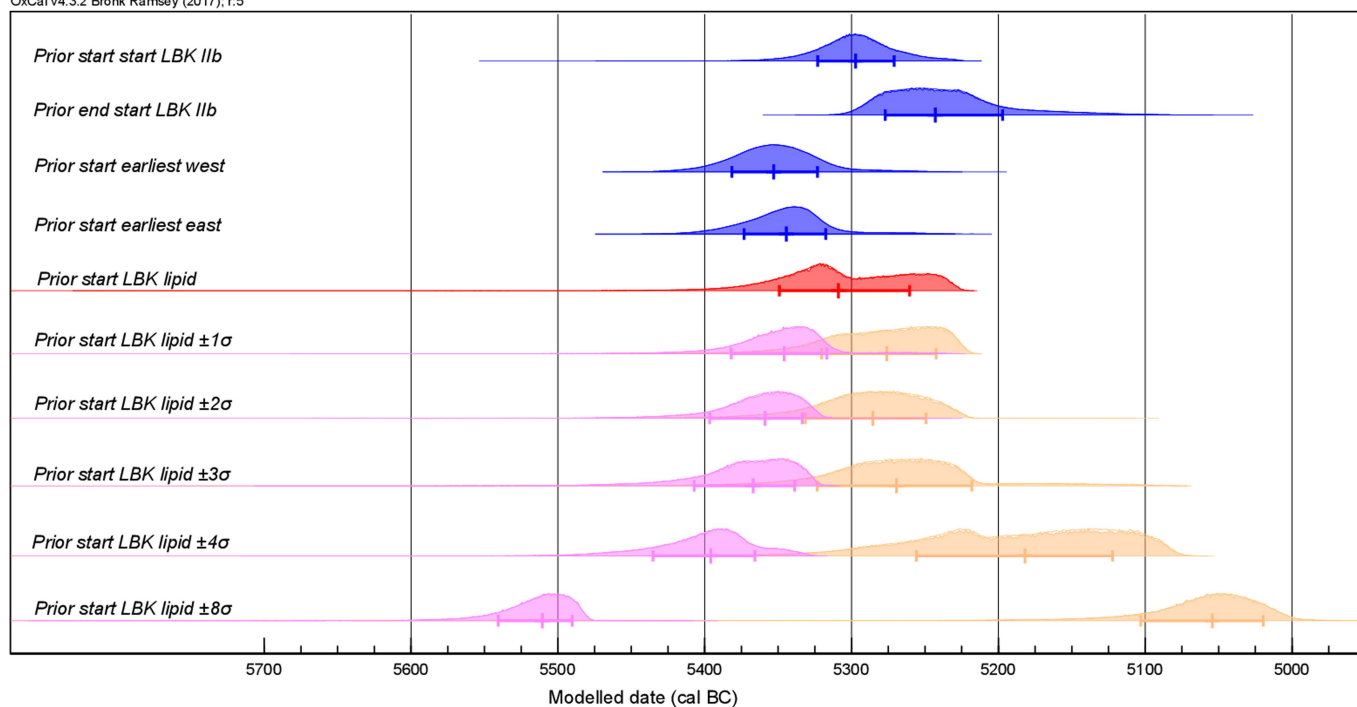


Extended Data Fig. 2 | Probability distributions of dates from Neolithic deposits in the TP area at Çatalhöyük, Turkey. Data include the results on absorbed fatty acids in pottery sherds listed in Extended Data Table 1. Each distribution represents the relative probability that an event occurs at a particular time. For each date, two distributions are plotted: one in outline, which is the result of a simple radiocarbon calibration, and a solid one, based on the chronological model used. The distributions in green correspond to the potsherds, distributions in black show the pre-existing chronology.

Distributions other than those relating to particular samples correspond to aspects of the model. For example, the distribution 'end East Mound occupation' is the estimated date at which the Neolithic occupation of the East Mound ended at Çatalhöyük. Measurements followed by a question mark and shown in outline have been excluded from the model for reasons described in table 1 of a previous study¹¹ and are simple calibrated dates³⁴. The large square brackets down the left side, along with the OxCal keywords, define the overall model exactly.

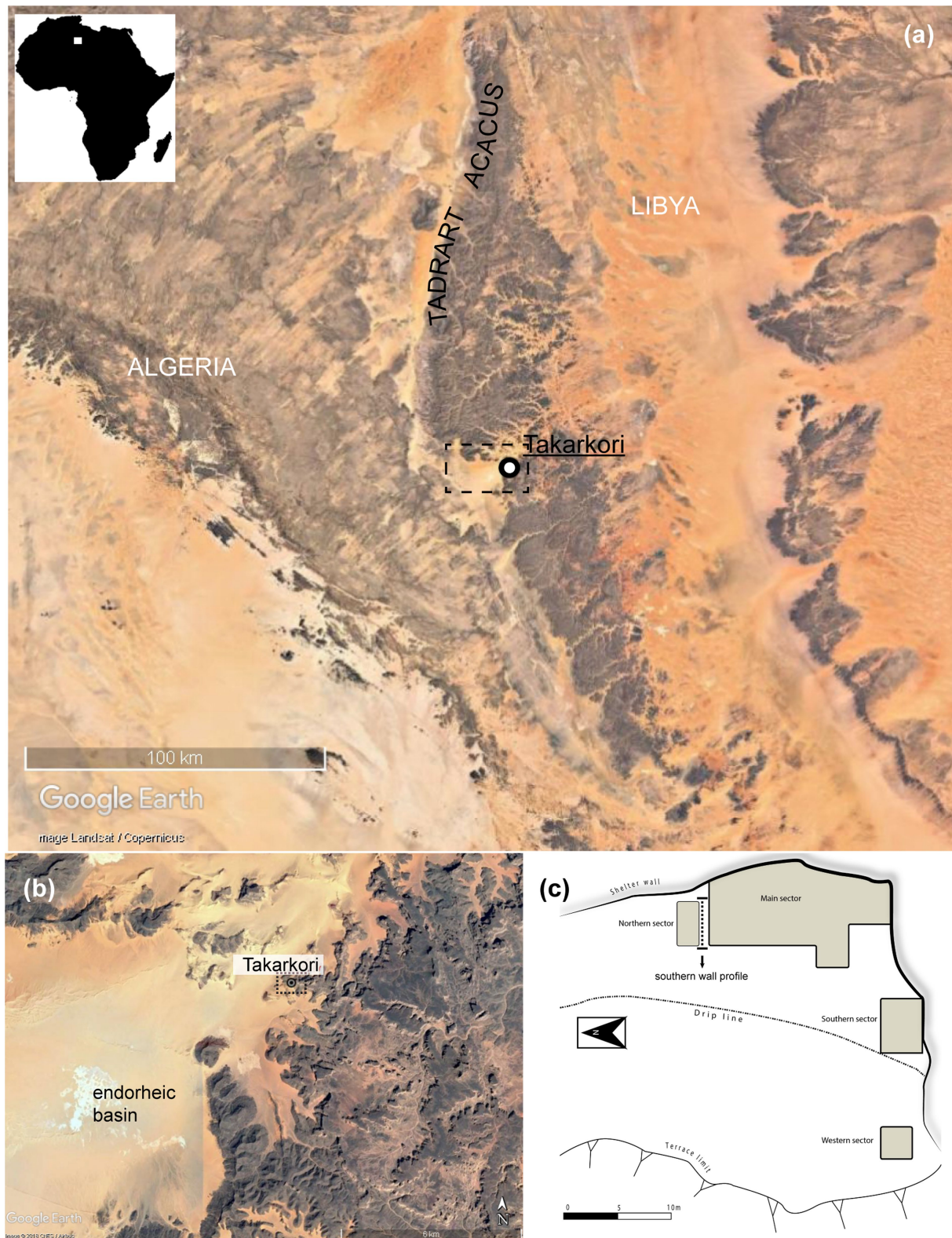


Extended Data Fig. 3 | Probability distributions of radiocarbon dates from absorbed fatty acids in LBK ceramics. Data on absorbed fatty acids are listed in Extended Data Table 1. Black, dairy; blue, ruminant adipose; red, non-ruminant adipose. Data are shown as described for Extended Data Fig. 2.

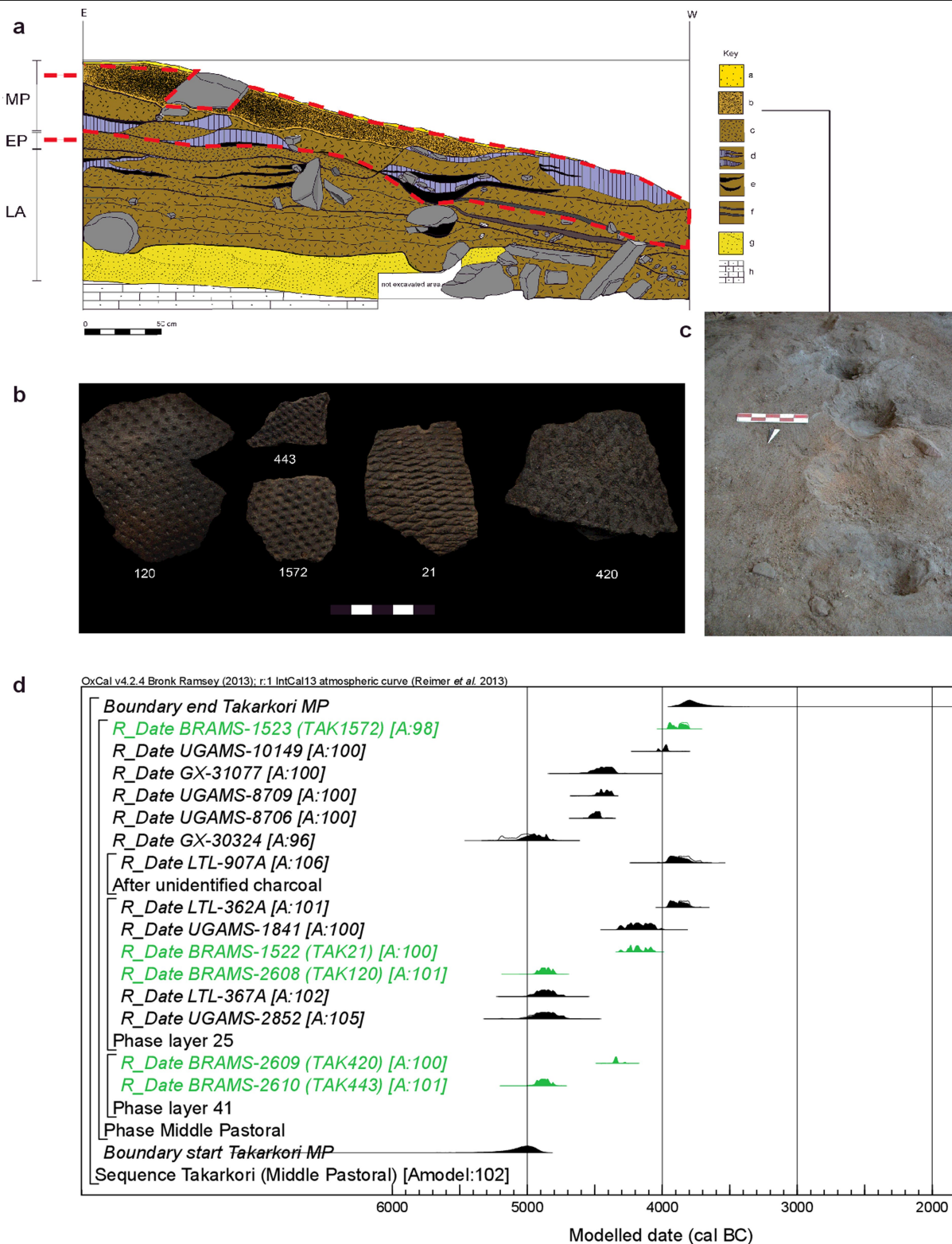


Extended Data Fig. 4 | Sensitivity analyses of radiocarbon dates on LBK ceramics. Key parameters for the start of the use of LBK ceramics (blue distribution)—derived from the models defined in Extended Data Fig. 3, figure 8 of a previous study¹², and figures 18, 19 (model 1), 20, 21 (model 2) and 22, 23 (model 3) of a previous publication¹³—were compared with the start of LBK

lipids presented in Extended Data Fig. 3 (red distributions), and subsequently deliberately biased by 1σ , 2σ , 3σ , 4σ and 8σ to younger (orange distributions) and older (pink distributions) values. Some distributions may have been truncated.

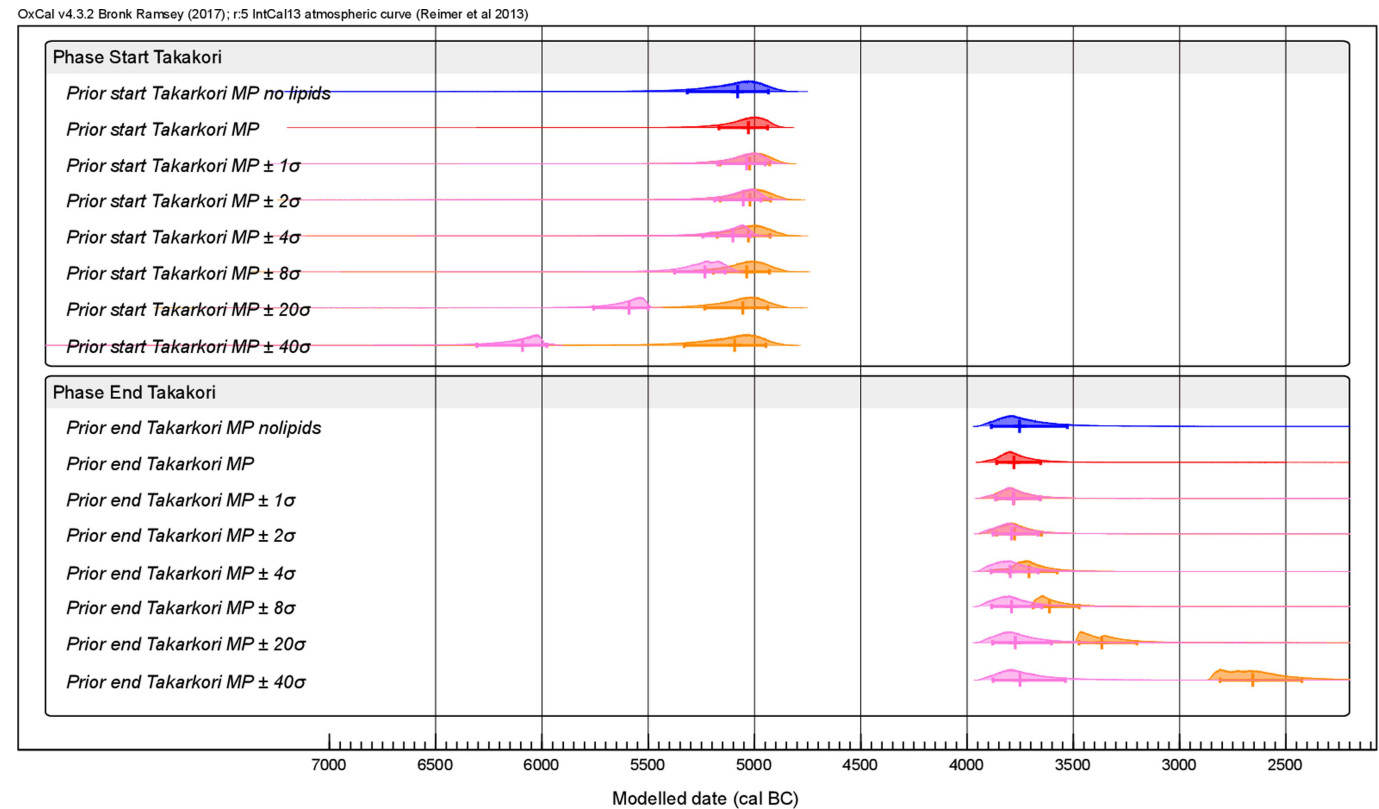


Extended Data Fig. 5 | The Tadrart Acacus Mountains in southwest Libya. a, b, The Wadi Takarkori area (dashed rectangle). c, Schematic plan of the excavated areas. All sampled sherds come from the main sector.

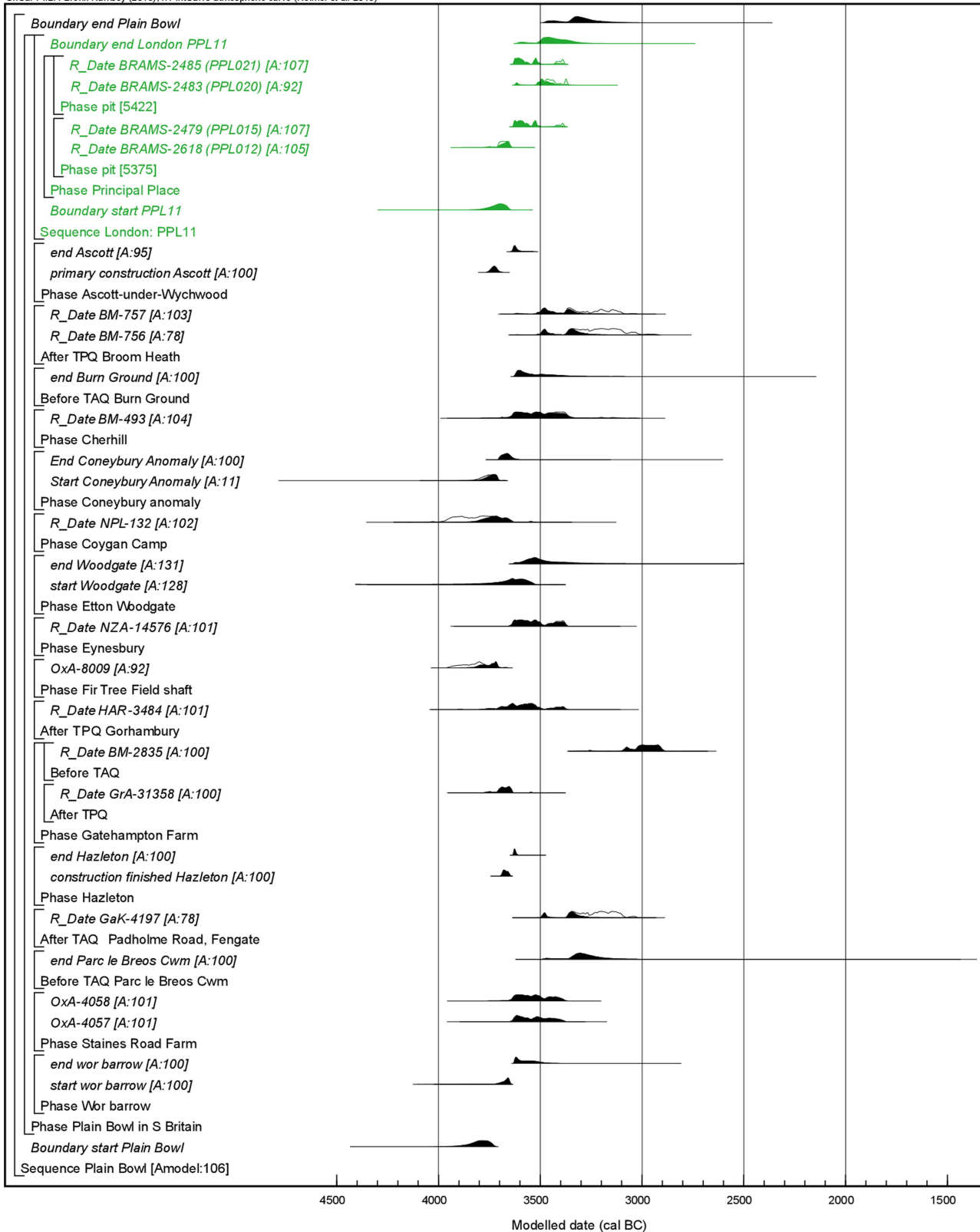


Extended Data Fig. 6 | Site stratigraphy, photographs of potsherds and radiocarbon dates of Middle Pastoral pottery vessels from Takarkori (Libya) modelled using Bayesian statistics. **a**, Stratigraphic context of sampled potsherds from Takarkori east–west profile of the southern wall of the Takarkori north–south (Extended Data Fig. 5). (a) aeolian sand; (b) sand rich in organic matter; (c) lenses of undecomposed plant remains; (d) ash; (e) charcoal; (f) slurry deposit; (g) eroded sand from the wall; (h) bedrock. **b**, Photographs of

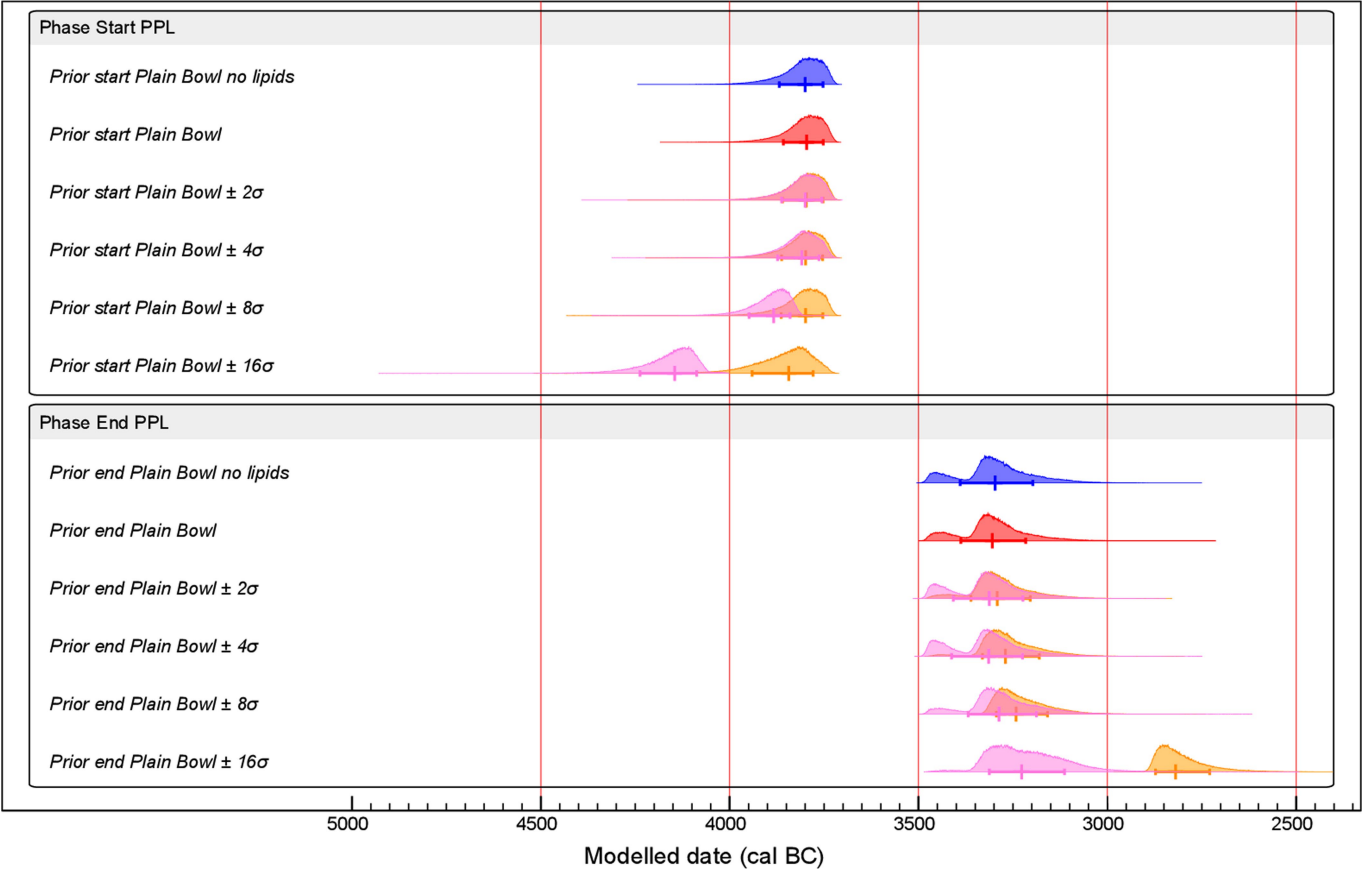
the five potsherds analysed showing typical Middle Pastoral decorative patterns. **c**, Example of temporally and spatially wide deposit of organic sands (detail of layer 25, Takarkori main sector). **d**, Statistical model of the Middle Pastoral period showing the comparison of pot lipid dates (in green) with previous radiocarbon measurements. Data are shown as described for Extended Data Fig. 2.



Extended Data Fig. 7 | Sensitivity analyses of radiocarbon dates on vessels from Takarkori rock shelter, Libya. Probability distributions for the beginning and end Middle Pastoral period ceramics from Takarkori rock shelter, Libya (no pot lipid dates) compared with those of the model shown in the Extended Data Fig. 6d and models including fatty acid dates that were deliberately biased by 1σ , 2σ , 4σ , 8σ , 20σ and 40σ . Data are shown as described for Extended Data Fig. 4.



Extended Data Fig. 8 | Probability distributions of dates associated with the use of early Neolithic Plain Bowl pottery in southern Britain. Prior distributions have been taken from the models described in the text and in the Supplementary Information. Data are shown as described for Extended Data Fig. 2.



Extended Data Fig. 9 | Sensitivity analyses of radiocarbon dates on vessels from Principal Place, London. Probability distributions of the start and end of early Neolithic Plain Bowl pottery in southern Britain compared with those of the model shown in Extended Data Fig. 8 and models including fatty acid dates that were deliberately biased by 2σ , 4σ , 8σ and 16σ . Data are shown as described for Extended Data Fig. 4.

Extended Data Table 1 | Summary of radiocarbon dates of lipids preserved in pottery vessels

Site	Location	Potsherd #	Description	C ⁹ (µg/g)	Laboratory#	C _{16:0} age	C _{18:0} age	Combined age	Reference
Sweet Track	Somerset levels, England	SW1	Carinated Bowl, refitted sherd	13,806	BRAMS-1520	5,105 ± 33	5,114 ± 32	5,110 ± 25	9, 10, 25
		SW2	Carinated Bowl, refitted sherd	4,900	BRAMS-1521	5,089 ± 38	5,094 ± 32	5,092 ± 26	
Çatalhöyük East 'TP Area'	Konya, Turkey	TP.M17	Holemouth jar, single sherd	393	BRAMS-1654	7,338 ± 42	7,416 ± 39	7,382 ± 31	11, 27
		TP.N10	Holemouth jar, refitted sherd	575	BRAMS-1699	7,318 ± 29	7,378 ± 30	7,348 ± 25	
		TP.O23	Holemouth jar, refitted sherd	1,390	BRAMS-1546	7,290 ± 36	7,375 ± 32	7,340 ± 27	
		TP.P13	Holemouth jar, single sherd	362	BRAMS-1703	7,328 ± 36	7,394 ± 29	7,364 ± 25	
Rosheim 'Sandgrube'	Lower Alsace, France	ROS-C-4596	Coarse Kumpf, single sherd	973	BRAMS-1526	5,810 ± 30	5,798 ± 30	5,804 ± 25	12
		ROS-C-4600	Coarse Kumpf, refitted sherd	4,163	BRAMS-1527	5,897 ± 36	5,909 ± 35	5,904 ± 28	
		ROS-C-4644	Fine Kumpf, single sherd	6,064	BRAMS-1525	5,937 ± 33	5,926 ± 30	5,931 ± 26	
		ROS-C-4657	Coarse Kumpf, single sherd	1,914	BRAMS-1524	5,885 ± 37	5,934 ± 34	5,912 ± 28	
Ensisheim 'Ratfeld'	Upper Alsace, France	ENS-C-5913	Coarse Kumpf, single sherd	1,177	BRAMS-1915	6,345 ± 31	6,303 ± 31	6,324 ± 26	12, 13
		ENS-C-5915	Coarse Kumpf, single sherd	771	BRAMS-1916	6,383 ± 32	6,314 ± 33	6,348 ± 26	
		ENS-C-5934	Coarse Kumpf, single sherd	1,645	BRAMS-1958	6,282 ± 30	6,258 ± 30	6,270 ± 25	
		ENS-C-5940	Coarse Kumpf, single sherd	2,082	BRAMS-2031	6,162 ± 33	6,239 ± 30	6,206 ± 26	
Cuiry-lès-Chaudardes	Aisne, France	CUI-C-5708	Coarse Kumpf, single sherd	881	BRAMS-1917	6,252 ± 34	6,218 ± 36	6,236 ± 27	12, 13
		CUI-C-5801	Coarse Kumpf, single sherd	9,886	BRAMS-2021	6,138 ± 30	6,134 ± 30	6,136 ± 25	
Königshoven 14	Rhineland, Germany	KON-C-5594	Coarse Kumpf, single sherd	531	BRAMS-2029	6,253 ± 29	6,298 ± 29	6,276 ± 24	12, 13
		KON-C-5598	Coarse Kumpf, single sherd	1,023	BRAMS-2026	6,106 ± 34	6,139 ± 34	6,123 ± 27	
Geleen-Janskamperveld	Graetheide, The Netherlands	GEL-C-3298	Coarse Kumpf, single sherd	577	BRAMS-2032	6,188 ± 31	6,253 ± 29	6,224 ± 25	12, 13
Karwowo 1	Pomerania, Poland	KAR-C-3636	Coarse Kumpf, single sherd	3,316	BRAMS-2028	6,176 ± 30	6,230 ± 30	6,204 ± 25	12, 13
		KAR-C-3677	Coarse Kumpf, single sherd	1,900	BRAMS-2025	6,255 ± 30	6,214 ± 32	6,236 ± 26	
Ludwinowo 7	Kuyavia, Poland	LDW-C-2267	Coarse Kumpf, single sherd	323	BRAMS-2024	6,173 ± 36	6,179 ± 30	6,177 ± 26	12, 13
Takarkori Rockshelter	Acacus mountains, Libya	TAK 21	Decorated, single sherd	9,503	BRAMS-1522	5,362 ± 33	5,331 ± 32	5,348 ± 24	14, 28, 29
		TAK1572	Decorated, single sherd	3,558	BRAMS-1523	5,099 ± 38	5,071 ± 32	5,085 ± 24	
		TAK 120	Decorated, single sherd	5,593	BRAMS-2608	6,008 ± 35	5,949 ± 35	5,979 ± 28	
		TAK 420	Decorated, single sherd	1,119	BRAMS-2609	5,487 ± 34	5,498 ± 35	5,493 ± 28	
		TAK 443	Decorated, single sherd	17,217	BRAMS-2610	6,021 ± 35	5,962 ± 35	5,992 ± 28	
Principal Place	London, England	PPL012 (<1814>)	Plain Bowl, single sherd	713	BRAMS-2618	4,894 ± 34	4,928 ± 33	4,911 ± 27	15
		PPL015 (<1845>)	Plain Bowl, refitted sherd	1,999	BRAMS-2479	4,708 ± 33	4,771 ± 30	4,742 ± 22	
		PPL020 (<1850>)	Plain Bowl, refitted sherd	3,660	BRAMS-2483	4,628 ± 40	4,670 ± 34	4,652 ± 26	
		PPL021 (<1819>)	Plain cup, single sherd	2,985	BRAMS-2485	4,732 ± 32	4,734 ± 30	4,733 ± 22	

Vessel descriptions, lipid concentrations and conventional radiocarbon ages (as defined previously³¹ and calculated according to previously published methods³⁰) of C_{16:0} and C_{18:0} fatty acids (which passed the internal quality control) extracted from pottery vessels.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☐ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☐ ☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data (radiocarbon dates) were acquired on a MICADAS AMS system using BATS software (v4.07)

Data analysis

The software BATS (v4.07) was used for data reduction analysis of radiocarbon dates and OxCal (v4.2 and v4.3) for chronological modeling

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data generated during this study are included in the main article, extended data and supplementary information.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

nature research | reporting summary

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The vessels dated were selected based on their lipid concentration which determined the number of samples radiocarbon dated in this paper
Data exclusions	Radiocarbon dates that did not pass the internal criterion (chi-square test on C16 and C18 fatty acids radiocarbon ages) explained in the SI document were excluded from chronological modeling
Replication	No replication of radiocarbon dates was undertaken due to the destructive nature of the radiocarbon dating technique, and limited sample availability. Tests of repeatability have been presented in a previously published paper Casanova et al. (2018) cited in this paper.
Randomization	Randomization was not relevant to the study. Our aim was to check the accuracy of radiocarbon dates determined for pottery vessels from secure and well defined archaeological contexts.
Blinding	Blinding was not relevant to the study, see as above.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems		Methods	
n/a	Involved in the study	n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies	<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines	<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology	<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants		
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data		

The gut–brain axis mediates sugar preference

<https://doi.org/10.1038/s41586-020-2199-7>

Received: 12 April 2019

Accepted: 21 February 2020

Published online: 15 April 2020

 Check for updates

Hwei-Ee Tan^{1,2,4}, Alexander C. Sisti^{1,3,4}, Hao Jin^{1,3}, Martin Vignovich^{1,3}, Miguel Villavicencio^{1,3}, Katherine S. Tsang^{1,3}, Yossef Goffer³ & Charles S. Zuker^{1,3}✉

The taste of sugar is one of the most basic sensory percepts for humans and other animals. Animals can develop a strong preference for sugar even if they lack sweet taste receptors, indicating a mechanism independent of taste^{1–3}. Here we examined the neural basis for sugar preference and demonstrate that a population of neurons in the vagal ganglia and brainstem are activated via the gut–brain axis to create preference for sugar. These neurons are stimulated in response to sugar but not artificial sweeteners, and are activated by direct delivery of sugar to the gut. Using functional imaging we monitored activity of the gut–brain axis, and identified the vagal neurons activated by intestinal delivery of glucose. Next, we engineered mice in which synaptic activity in this gut-to-brain circuit was genetically silenced, and prevented the development of behavioural preference for sugar. Moreover, we show that co-opting this circuit by chemogenetic activation can create preferences to otherwise less-preferred stimuli. Together, these findings reveal a gut-to-brain post-ingestive sugar-sensing pathway critical for the development of sugar preference. In addition, they explain the neural basis for differences in the behavioural effects of sweeteners versus sugar, and uncover an essential circuit underlying the highly appetitive effects of sugar.

Sugar is a fundamental source of energy for all animals, and correspondingly, most species have evolved dedicated brain circuits to seek, recognize and motivate its consumption⁴. In humans, the recruitment of these circuits for reward and pleasure—rather than nutritional needs—is thought to be an important contributor to the overconsumption of sugar and the concomitant increase in obesity rates. In the 1800s the average American consumed less than 4.5 kg of sugar per year⁵; today, following the broad availability of refined sugar in consumer products, the average consumption is more than 45 kg per year⁶.

Sweet compounds are detected by specific taste receptor cells on the tongue and palate epithelium^{7,8}. Activation of sweet taste receptor cells sends hardwired signals to the brain to elicit recognition of sweet-tasting compounds^{9,10}. We and others have studied the circuits linking activation of sweet taste receptors on the tongue to sweet-evoked attraction^{8,11,12}. Surprisingly, even in the absence of a functional sweet-taste pathway, animals can still acquire a preference for sugar^{1,2,7}. Furthermore, although artificial sweeteners activate the same sweet taste receptor as sugars, and they may do so with vastly higher affinities⁷, they fail to substitute for sugar in generating a behavioural preference¹³.

Together, these results suggested the existence of a sugar-specific, rather than sweet-taste-specific pathway, that operates independently of the sense of taste to create preference for sugar and motivate consumption^{2,14}. Here, we dissect the neural basis for sugar preference.

Sweet versus sugar preference

When non-thirsty, wild-type mice are given a choice between water and sugar they drink almost exclusively from the sugar solution⁷. If, however, they are allowed to choose between an artificial sweetener (for example, acesulfame K (AceK)) and sugar, using concentrations at which both are equally attractive, naive mice initially drink from both bottles at a similar rate (Fig. 1a). However, within 24 h of exposure to both choices, their preference is markedly altered, such that by 48 h, they drink almost exclusively from the bottle containing sugar (Fig. 1a, b, compare 15 h with 48 h). This behavioural switch also happens in knockout (KO) mice lacking sweet taste (*Trpm5*^{−/−} (hereafter TRPM5 KO)^{15,16} or *Tas1r2*^{−/−}*Tas1r3*^{−/−} (hereafter T1R2/3 KO)⁷; Fig. 1c). Similar observations have been made in several studies, primarily using flavour-conditioning assays^{1,2}. Thus, although taste-knockout mice cannot taste sugar or sweetener, they learn to recognize and choose the sugar, most probably as a result of strong positive post-ingestive effects¹⁷.

Notably, the preference for sugar does not rely on its caloric content¹⁸. For example, if sugar is substituted for the non-metabolizable glucose analogue (methyl- α -D-glucopyranoside (MDG))¹⁹ mice still develop a strong preference for MDG, just as they do for glucose (Fig. 1b; Extended Data Fig. 1). Thus, the signalling system recognizes the sugar molecule itself rather than its caloric content or metabolic products.

¹Zuckerman Mind Brain Behavior Institute, Howard Hughes Medical Institute and Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. ²Department of Biological Sciences, Columbia University, New York, NY, USA. ³Department of Neuroscience, Vagelos College of Physicians and Surgeons, Columbia University, New York, NY, USA. ⁴These authors contributed equally: Hwei-Ee Tan, Alexander C. Sisti. ✉e-mail: cz2195@columbia.edu

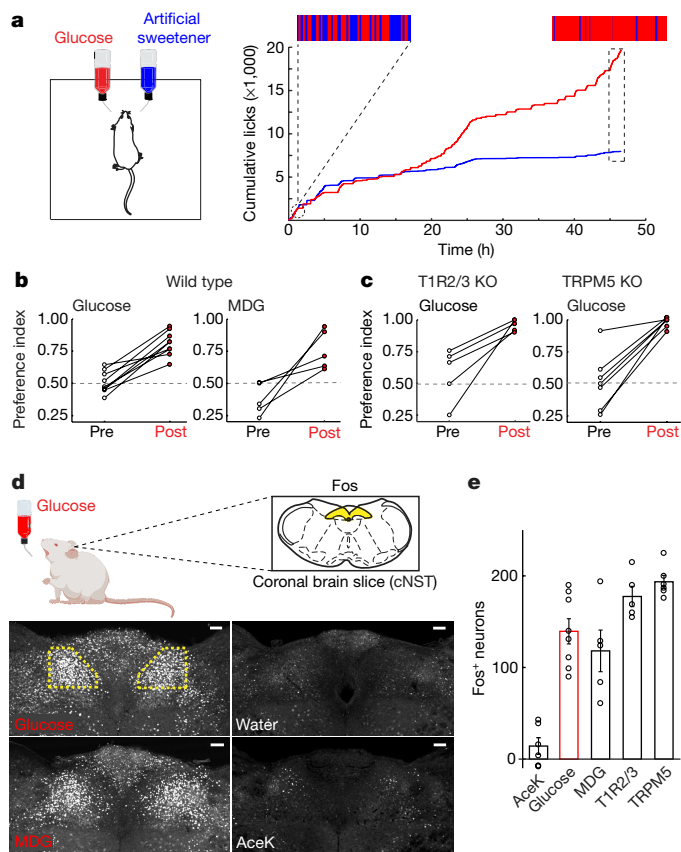


Fig. 1 | Sugar activates the gut–brain axis. **a**, Mice were allowed to choose between 600 mM glucose and 30 mM AceK (sweetener). Preference was tracked by electronic lick counters in each port. Bars at the top show lick rasters for glucose (red) versus AceK (blue) from the first and last 2,000 licks of the behavioural test. Note that by 48 h the animals drink almost exclusively from the sugar bottle. **b**, Preference plots for sugar versus AceK ($n = 9$ mice, two-tailed paired t -test, $P = 2.39 \times 10^{-6}$) and MDG versus AceK ($n = 5$ mice, two-tailed paired t -test, $P = 0.0024$; Extended Data Fig. 1). Note that mice may begin the behavioural preference test exhibiting no preference for sugar (preference index ≈ 0.5), some preference for sugar (preference index > 0.5) or with an initial preference for the sweetener (preference index < 0.5). However, in all cases they switched (or substantially increased) their preference towards sugar. **c**, Mice lacking the sweet taste receptor (T1R2/3 KO)⁷ ($n = 5$ mice, two-tailed paired t -test, $P = 0.0038$), and mice lacking TRPM5 (TRPM5 KO)¹⁵ ($n = 7$ mice, two-tailed paired t -test, $P = 0.0001$) switched their preference to sugar even though they cannot taste it. **d**, Schematic of sugar stimulation of Fos induction. Strong Fos labelling is observed in neurons of the cNST (highlighted yellow). Scale bars, 100 μ m. Similar results were obtained in multiple mice in each experiment (Extended Data Fig. 2). **e**, Quantification of Fos-positive neurons. The equivalent area of the cNST (bregma -7.5 mm) was processed and counted for the different stimuli. The signal present in water alone was subtracted before plotting; ANOVA with Tukey's honestly significant difference (HSD) post hoc test against AceK ($n = 6$ mice): $P = 4.68 \times 10^{-5}$ (glucose, $n = 8$ mice), $P = 0.001$ (MDG, $n = 5$ mice). Values are mean \pm s.e.m.

Brain neurons activated by sugar

For an animal to develop a preference for sugar over sweetener, it must recognize and distinguish between two innately attractive stimuli. We reasoned that if we could identify a population of neurons that respond selectively to the consumption of sugar, it may provide an entry to reveal the neural control of sugar preference and the basis of sugar craving.

We exposed separate cohorts of mice to sugar, sweetener or water, and examined their brains for induction of Fos as a proxy for neural activity²⁰ (see Methods). Our results showed prominent bilateral

labelling in the caudal nucleus of the solitary tract (cNST; Fig. 1d), an area known to function as a nexus of interoceptive signals conveying information from the body to the brain²¹. By contrast, the cNST was not substantially labelled in response to sweetener or water controls (Fig. 1d, Extended Data Fig. 2a). Furthermore, if these cNST neurons are involved in sugar-preference behaviour, they must also be activated by MDG (Fig. 1d, e), and their activation by sugar should be independent of the taste system (Extended Data Fig. 3a).

How do sugar signals reach the cNST? The finding that preference for sugar does not require the taste system strongly suggested post-ingestive recognition. Therefore, we tested whether intragastric application of sugar was sufficient to activate the cNST. As predicted, direct gut infusion of sugar (but not sweetener) is sufficient to activate the cNST as robustly as oral ingestion (Extended Data Fig. 3b). These results also substantiate previous behavioural studies showing that intragastric infusion of glucose is sufficient to condition flavour preference^{22,23}.

The gut–brain axis

A number of recent studies have implicated the gut–brain axis as a key mechanism for transmitting information from the gut to the brain via the vagus nerve^{24–26}. The gut–brain axis is emerging as a fundamental conduit for the transfer of neural signals informing the brain of the metabolic and physiologic state of the body. If information about sugar detection is being transferred from the gut to the cNST via the gut–brain axis, then it should be possible to directly monitor the activity of this circuit by using real-time recordings of cNST activation in response to synchronized gut stimulation with sugar. Furthermore, this activity should be abolished following transection of the vagal nerve, and notably, silencing vagal sensory neurons should prevent the creation of sugar preference.

We used fibre photometry²⁷ to record sugar-evoked responses in the cNST of mice expressing the genetically encoded calcium indicator GCaMP6s in excitatory neurons (*Vglut2-cre*;Ai96; *Vglut2* is also known as *Slc17a6*). To deliver stimuli to the gut, we placed a catheter directly into the duodenal bulb and created an exit port by transecting the intestine about 12 cm distally (Fig. 2a, see Methods). As predicted, our results (Fig. 2b–d) showed robust responses to glucose and MDG. Most notably, all activity was abolished after bilateral transection of the vagal nerves (Fig. 2b–e).

Next, we examined whether cNST neurons activated in response to sugar ingestion indeed receive direct input from vagal ganglion neurons (that is, from the nodose ganglia; see Extended Data Fig. 4a, b). To test this, we used the targeted recombination in active populations (TRAP) system^{28,29} to target Cre recombinase to sugar-activated cNST neurons, and used a Cre-dependent monosynaptic retrograde viral reporter to identify their synaptically connected input neurons^{30,31}.

We infected the cNST with adeno-associated virus (AAV) carrying a Cre-dependent glycoprotein coat and a surface receptor for a trans-synaptic reporter virus^{30,31}, and TRAPed sugar-activated neurons (Fig. 3a; see Extended Data Fig. 4c, d for selectivity of TRAPing). Next, we infected the TRAPed neurons with the retrograde rabies reporter (RABV–dsRed), and investigated whether sugar-activated cNST neurons receive input from vagal ganglion neurons. As controls, we carried out similar experiments but used water or sweetener as TRAPing stimuli. Our results (Fig. 3b, c) revealed large numbers of nodose neurons labelled by the transsynaptic tracing strategy, demonstrating that the sugar-activated cNST neurons receive direct monosynaptic input from the vagal ganglion. By contrast, when we used AceK or water for TRAPing, only a small number of vagal neurons were labelled; we believe these represent activation to licking/drinking or ingestion (Fig. 3b, c).

Finally, we carried out a genetic vagotomy by globally silencing nodose sensory neurons (see Methods), which—as predicted—prevented the development of sugar preference (Extended Data Fig. 3c, d).

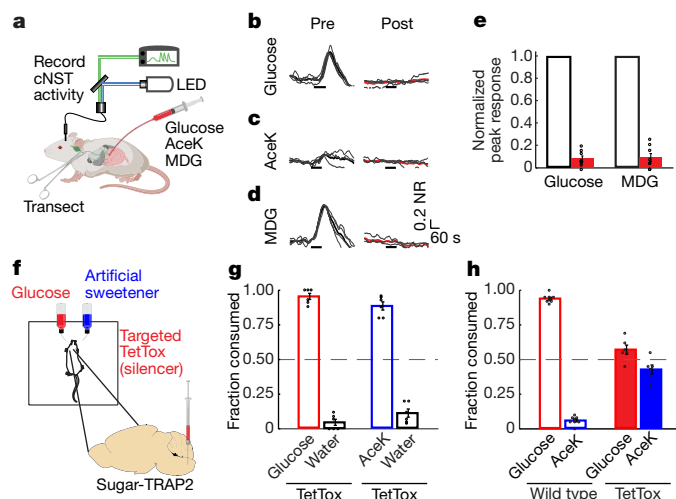


Fig. 2 | Silencing the sugar-activated circuit abolishes sugar preference. **a**, Fibre photometry monitoring glucose-evoked responses of cNST neurons. The excitatory neurons in the cNST were targeted with GCaMP6s using *VGlut2-cre* mice⁴⁶. **b–d**, Neural responses following intestinal delivery of glucose, AceK or MDG. Note strong responses to sugar (**b**) and MDG (**d**). The light traces denote normalized two-trial averages from individual animals and the dark trace is the average of all trials. Black bars below traces indicate the time and duration of stimuli. The average responses after bilateral vagotomy are shown in red (see Methods). Stimuli: 500 mM glucose, 30 mM AceK or 500 mM MDG; $n = 4$ mice. NR, normalized response. **e**, Quantification of neural responses before and after vagotomy. Two-tailed paired *t*-test, $P = 3 \times 10^{-15}$ (glucose), $P = 5 \times 10^{-13}$ (MDG), $n = 4$ mice. Data are mean \pm s.e.m. **f**, Schematic of silencing strategy. TRAP2 mice²⁹ were stimulated with 600 mM glucose to induce expression of Cre recombinase in the cNST. AAV-DIO-TetTox³² was then targeted bilaterally to the cNST for silencing. **g**, Silencing the sugar-preference neurons in the cNST does not impair the innate attraction to sugar or sweeteners. The graph shows preference for 600 mM glucose versus water, and preference for 30 mM AceK versus water. $n = 6$ mice. Data are mean \pm s.e.m. **h**, Sugar-preference graphs (48-h tests) for wild-type mice, demonstrating the robust development of preference for sugar versus sweetener (see also Fig. 1). By contrast, silencing sugar-activated neurons in the cNST abolishes the development of sugar preference. $n = 6$ mice; two-sided Mann–Whitney *U*-test, $P = 4 \times 10^{-4}$; TetTox-silenced animals consumed as much of the AceK sweetener as they did sugar (see also Extended Data Fig. 5). Data are mean \pm s.e.m.

Neurons in the cNST mediate sugar preference

If the gut-to-brain sugar-activated cNST neurons are essential for creating preference for sugar, then blocking their function should prevent the formation of sugar preference. Therefore, we engineered mice in which synaptic transmission in the sugar-preference neurons was genetically silenced by targeted expression of tetanus toxin light chain (TetTox)³². Our strategy relied on the TRAP system^{28,29} to express inducible Cre recombinase in sugar-activated cNST neurons, and bilaterally injecting the cNST with an AAV carrying the Cre-dependent TetTox construct (Fig. 2f, see Methods).

First, we needed to ensure that silencing this circuit did not affect the innate ability of the animals to be attracted to sweet taste, including sugar and sweeteners. Indeed, when the TetTox-targeted mice were tested to choose between sweet or water, they selected the sweet-tasting solutions (either AceK or glucose; Fig. 2g). However, silencing the sugar-activated cNST neurons abolished their capacity to develop a preference for sugar over artificial sweetener, even after prolonged testing sessions (Fig. 2h, Extended Data Fig. 5). These results illustrate the essential role of this circuit in driving the behavioural preference for sugar.

Vagal neurons sensing gut sugar

As information about sugar detection is being transferred via the gut–vagal–brain axis, we set out to directly monitor the activity of

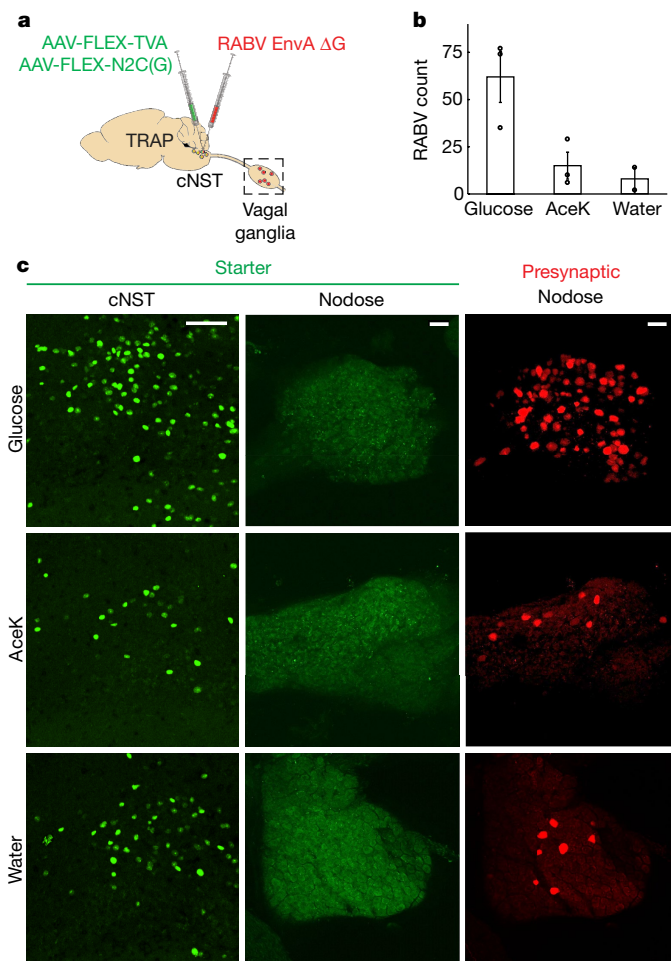


Fig. 3 | Vagal ganglion neurons transmit sugar signals to the brain. **a**, Strategy for targeting a red fluorescently labelled retrograde transsynaptic rabies reporter (RABV–dsRed)^{28,31} to the cNST. Sugar-TRAP neurons in the cNST (designated as ‘starter cells’)^{30,31} were infected with AAVs encoding proteins required for infection with RABV–dsRed, resulting in labelling of the monosynaptic inputs of the sugar-activated cNST neurons by the retrogradely transsynaptic transfer of the RABV–dsRed virus. **b**, Quantification of retrogradely labelled RABV–dsRed neurons in the nodose ganglion. Sugar versus AceK TRAP labelling ($n = 3$ mice). ANOVA, Tukey’s HSD post hoc test, $P = 0.0449$. We also performed control TRAP labelling with water ($n = 2$ animals). Sugar versus water TRAP: ANOVA, Tukey’s HSD post hoc test, $P = 0.0407$; AceK versus water TRAP: $P = 0.9$. Data are mean \pm s.e.m. **c**, Sugar-TRAP cNST neurons (starter, green) receive monosynaptic input from vagal neurons (RABV, red). Note the absence of starter cells in the nodose, confirming that the RABV (red) cells represent retrogradely labelled neurons^{30,31}. Scale bars, 100 μ m.

this circuit by imaging vagal-neuron responses to gut stimulation with sugar.

We implemented a vagal ganglion functional imaging platform (Fig. 4a) by targeting the genetically encoded calcium indicator GCaMP³³ to vagal sensory neurons³⁴ (*Vglut2-cre*;Ai96). To visualize the neurons and measure calcium dynamics, we exposed a 1-cm² ventral window into the ganglion and used a one-photon microscope equipped with an electron-multiplying CCD camera for imaging³⁵. For most imaging sessions, the intestinal segment was exposed to a pre-stimulus application of PBS, then a 10-s (33 μ l) or 60-s (200 μ l) test stimulus, and a 3-min post-stimulus wash. Neuronal signals were analysed for statistically significant increases in intracellular calcium over baseline (see Methods), and a neuron was classified as a responder if it exhibited responses in more than 60% of the trials³⁶.

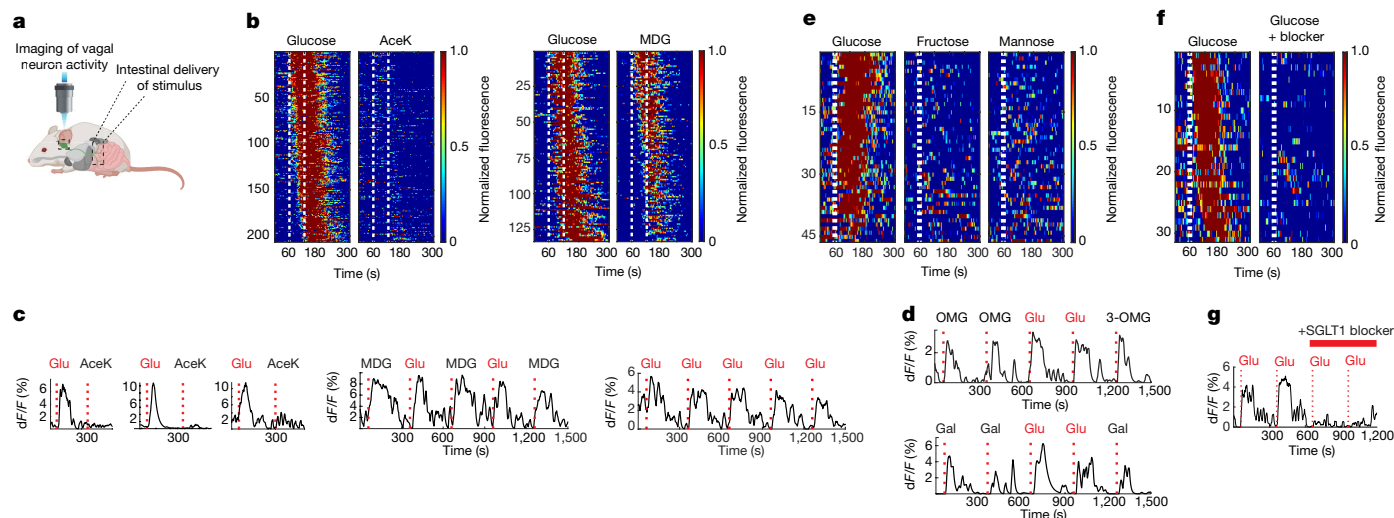


Fig. 4 | Imaging the gut–brain axis. **a**, We imaged calcium responses in vagal sensory neurons expressing the fluorescent calcium indicator GCaMP6s while stimulating the intestines. **b**, Heat maps depicting z-score-normalized fluorescence traces^{33,47} from vagal neurons identified as glucose responders. Each row represents the average activity of a single cell to three trials. Stimulus window is shown by dashed white lines. Left, responses of $n = 206$ vagal neurons to a 60-s intestinal infusion of 500 mM glucose; note lack of responses to 30 mM AceK. Right, heat maps depicting $n = 133$ vagal neurons that responded to 60-s infusion of 500 mM glucose, and tested for their responses to 500 mM MDG. Heat maps were normalized across stimuli; responses to glucose and MDG were similar (two-tailed paired t -test, $P = 0.06$). **c**, Sample traces of vagal-neuron responses to intestinal stimulation with 60-s pulses of 30 mM AceK and 500 mM glucose from 3 mice (top), or to 10-s pulses of 500 mM glucose and 500 mM MDG (middle and bottom). Note the reliability and rapid onset of responses to the 10-s stimulus (Extended Data Fig. 6c). When using a 10-s stimulus, to minimize potential osmolarity responses (Extended Data

Fig. 8), approximately 5% of imaged neurons show statistically significant responses to glucose (Extended Data Fig. 6d). We compared imaging sessions with both the right and left ganglia²⁵ and did not observe any meaningful difference in the proportion of glucose-responding neurons (Extended Data Fig. 6e). **d**, Vagal-neuron responses to 3-OMG (top) and galactose (bottom), $n = 3$ independent experiments each. These agonists activate vagal neurons in a similar manner to glucose (Extended Data Figs. 2b, 10a, b). **e**, Heat maps of 46 glucose-responding neurons to 500 mM fructose and 500 mM mannose ($n = 5$ ganglia). The monosaccharides fructose and mannose, which are not substrates for SGLT1, do not activate glucose-responsive neurons. Fewer than 10% of glucose responders were activated by either fructose or mannose. **f, g**, Summary of responses to a 10-s stimulus of 500 mM glucose for 33 neurons before and after intestinal application 8 mM phlorizin for 5 min ($n = 4$ mice). Responses are severely diminished after blocker application (see Extended Data Fig. 10d, e and Methods).

First, we examined how vagal neurons respond to intestinal delivery of glucose versus sweetener. Delivering glucose into the intestines elicited significant calcium responses in subsets of ganglion neurons (Fig. 4b); we analysed the responses from the vagal ganglia of 8 different mice to a 60-s stimulus of glucose or AceK, and identified around 200 neurons that displayed statistically significant responses to glucose, but less than 1% of these neurons displayed stimulus-dependent activity to AceK (Fig. 4b). As expected, intestinal delivery of MDG also activated the majority of vagal neurons that responded to glucose (Fig. 4b, c, Extended Data Fig. 6a).

Next, we assessed the reliability and temporal causality of the vagal responses by reducing the stimulus window from 60 s to 10 s. Our results showed that vagal responses to intestinal glucose are robust and reliable³⁷ (Fig. 4c, Extended Data Fig. 6b, c). Overall, we examined 51 ganglia and 4–5% of GCaMP-expressing neurons (205 out of 4,803 neurons) responded to the 10-s glucose stimulus (Extended Data Fig. 6d).

As neurons in the nodose ganglion innervate the gut²¹ (that is, the source of the gut–brain signal), the cell bodies of the sugar-sensing neurons in the nodose ganglion should be retrogradely labelled by applying a tracer from their afferents in the gut²⁶. Thus, we injected fluorescently conjugated cholera toxin subunit B (CTB)³⁸ into the duodenum of GCaMP-expressing mice (Extended Data Fig. 7a), and examined the labelled duodenal innervating neurons for responses to intestinal delivery of sugar. Indeed, around 20% of the duodenum back-filled vagal sensory neurons robustly responded to glucose (Extended Data Fig. 7b, c).

We note that a previous study reported the characterization of candidate nutrient-sensing neurons in the nodose ganglia²⁴. These neurons responded indiscriminately to high concentrations of several

stimuli, including 1 M glucose and 0.5 M salt. Our results show that such responses, which are largely independent of the quality of the stimulus, are not glucose-sensing nor are they required for the development of sugar preference, but rather represent responses to a wide range of high-osmolarity stimuli (Extended Data Figs. 8, 9).

SGLT1 transduces gut–brain sugar signals

We reasoned that the gut-to-brain signal might depend on known sugar sensors recruited into this role, perhaps in a dedicated subpopulation of gut cells. Although the sweet-taste receptor is expressed in enteroendocrine cells³⁹, it is not involved in this process, as sweet-taste receptor knockout (T1R2/3 KO) mice still exhibited normal sugar-preference behaviour (Fig. 1c).

The principal glucose transporter (and sensor) in the intestine is the sodium–glucose-linked transporter-1 (SGLT1)^{19,40}. This transporter is expressed in enterocytes as well as in enteroendocrine cells that secrete a wide range of hormones and bioactive molecules and are thought to also function as a conduit between the gut and the vagal nerve^{41,42}. Therefore, we investigated whether SGLT1 is required to transmit the gut-to-brain sugar signal by determining whether other substrates of SGLT1—galactose and the glucose analogue 3-*O*-methyl-D-glucose (3-OMG)¹⁹—also activate the same vagal neurons as glucose. Indeed, neurons responding to intestinal glucose were also stimulated by 3-OMG and galactose (Fig. 4d, Extended Data Fig. 10a, b). Critically, this circuit is dedicated to glucose, as other caloric sugars such as fructose and mannose (that are not substrates of SGLT1)¹⁹ do not activate the glucose-responsive vagal neurons (Fig. 4e, Extended Data Fig. 10c), do not create a behavioural preference (Extended Data Fig. 8e), but still trigger osmolarity responses (Extended Data Fig. 8).

Next, we assessed whether pharmacological inhibition of SGLT1 abolishes the glucose-dependent neuronal responses. We examined the responses to two consecutive 10-s stimuli of intestinal glucose before and after a 5-min wash of the intestinal segment with phlorizin¹⁹, an SGLT1 blocker. Our results (Fig. 4f, g, Extended Data Fig. 10d, e), demonstrated a marked loss of glucose responses following intestinal application of phlorizin¹⁹. Together, these results place SGLT1 as an important component of the sugar-preference signalling circuit. It will be of interest to determine the identity of the intestinal cells mediating these responses, as they represent another potential target for modulating this circuit.

Co-opting the sugar-preference circuit

The results presented above reveal a specific circuit via the vagal ganglia to the brain critical for driving the development of preference for sugar. We devised an experiment to determine whether the selective activation of this circuit can be recruited to create a preference to a previously less-preferred stimulus. Our strategy was to identify a genetic driver that marks sugar-preference neurons in the cNST, and then link their activation to the ingestion of a novel stimulus.

We examined the Allen Brain Atlas for candidate genes with enriched expression in the cNST, and tested candidates for glucose-evoked Fos labelling (Fig. 5a). Our results demonstrated that proenkephalin (*Penk*)-expressing neurons in the cNST⁴³, marked by a *Penk-cre* construct driving tdTomato (*Penk-cre*;Ai75D), respond strongly to sugar stimuli (Fig. 5b, c); approximately 85% of the sugar-induced Fos-labelled neurons in the cNST are *Penk*-positive, and over 80% of the *Penk*-positive neurons were labelled by Fos after sugar ingestion.

We injected a Cre-dependent AAV encoding the excitatory designer receptor hM3Dq⁴⁴ into the cNST of *Penk-cre* mice, so that *Penk* cNST neurons could be experimentally activated by the hM3Dq agonists clozapine *N*-oxide or clozapine^{44,45}. After 8 days to allow for receptor expression, mice were exposed to two-bottle preference assays using artificially sweetened cherry-flavoured versus grape-flavoured solutions (Fig. 5d). Under this paradigm, the cherry solution was made sweeter than grape (see Methods) so that the animals would be significantly more attracted to the cherry flavour (Fig. 5e). Next, we introduced clozapine into the less-preferred grape flavour, and investigated whether clozapine-mediated activation of the *Penk* cNST neurons (much like glucose-mediated activation) can create a new preference. Indeed, after 48 h of exposure to the grape plus clozapine bottle, mice completely switched their preference, even though the grape solution was far less sweet than the cherry solution (Fig. 5e, purple lines). To demonstrate that the preference switch is independent of the nature of the initially less-preferred stimuli, we flipped the starting flavours so that cherry was less favoured, and obtained an equivalent switch in preference (Fig. 5e, red lines). As anticipated, wild-type mice without the designer receptor were indifferent to clozapine and continued to prefer the sweeter solution (Fig. 5f). These results demonstrate that artificial activation of the sugar-preference circuit is sufficient to drive the development of a novel preference to an otherwise low-preference stimulus.

Discussion

Sugar is an essential energy source across all animal species, and it is therefore expected that selective circuits be dedicated to seek, recognize and motivate its consumption. The discovery of this gut-to-brain circuit provides a powerful pathway to help to meet these needs.

In this study, we show that glucose acts in the gut to activate a neural circuit that communicates to the brain the presence of sugar. What is the advantage of a gut-to-brain sugar-detection system in addition to the taste system? A post-ingestive sensing system in the gut assures that signalling only occurs after the sugar molecules reach their desired

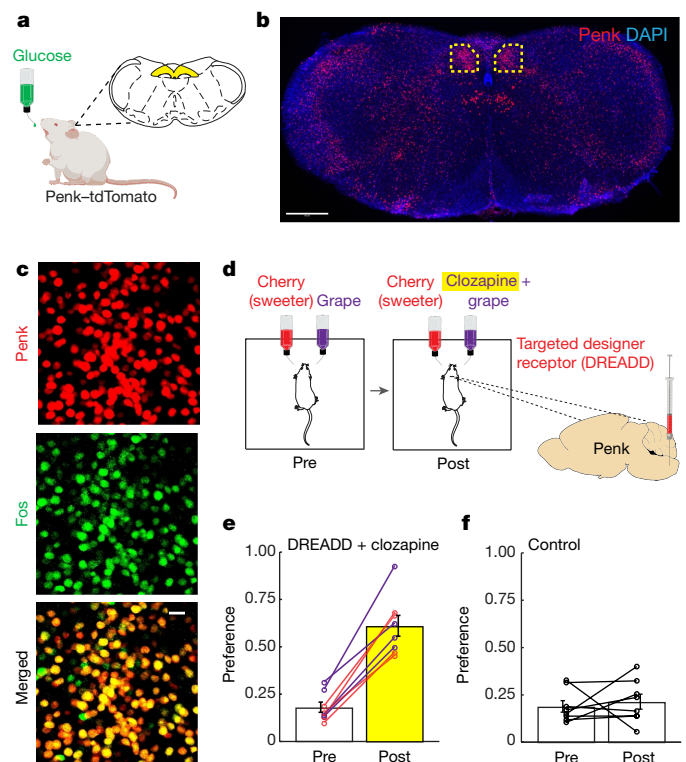


Fig. 5 | Activation of sugar-responsive cNST neurons confers novel flavour preference. **a–c**, *Penk-cre* mice were stimulated with 600 mM glucose and brain slices were analysed for Fos and *Penk* labelling. *Penk* neurons were marked by expression of nuclear-localized tdTomato (Ai75D reporter line)⁴⁸. **b**, Low-magnification section of the brain stem (bregma – 7.5 mm) showing *Penk* expression (red); tissue was counterstained with DAPI (blue). $n = 2$ independent experiments. Scale bar, 500 μ m; cNST, yellow box. **c**, Sugar-preference neurons express *Penk*. *Penk* neurons labelled with tdTomato (from **b**) and glucose-activated neurons (Fos-labelled) marked green. Note the high degree of overlap in the merged image. Approximately 85% of sugar-activated cNST neurons are marked by *Penk*, and about 90% of cNST *Penk* neurons show sugar-Fos labelling ($n = 3$ mice). Scale bar, 20 μ m. **d**, Expression of activating DREADD receptor^{44,45} (via AAV-DIO-hM3Dq) was targeted bilaterally to the cNST of *Penk-cre* mice. The mice were then tested for their preference between two flavours for 48 h (Pre). Shown is an example using cherry (containing 2 mM AceK) versus grape (with 1 mM AceK). Mice were conditioned and tested using the less-preferred flavour plus the DREADD agonist clozapine (Post; see Methods). **e**, *Penk*-hM3Dq mice initially prefer the sweeter solution. After associating clozapine-mediated activation of *Penk* cNST neurons with the less-preferred flavour, all the *Penk*-hM3Dq mice switched their preference (Pre, $18.1 \pm 2.7\%$; Post, $61.1 \pm 5.5\%$; $n = 8$ mice; two-sided Mann–Whitney *U*-test, $P = 1 \times 10^{-4}$). The experiment was carried out using grape (purple lines) or cherry (red lines) as the initially less-preferred stimuli. **f**, Mice not expressing the DREADD receptor are unaffected by the presence of clozapine (Pre, $19.0 \pm 3.0\%$; Post, $21.4 \pm 4.0\%$; $n = 8$ mice); control mice were subjected to the same conditioning and testing as the experimental cohort. Values are mean \pm s.e.m.

target for effective absorption and metabolic consumption. The association between the activation of this gut-to-brain circuit paired with the recognition of sugar by the taste system affords animals the fundamental capacity to identify, develop and reinforce a strong and durable preference for sugar-rich food sources. The evolutionary association of these two separate circuits combines nutrition with the basic sense of taste. In the future, it would be of interest to determine whether preference for other essential nutrients also utilizes this gut–brain axis.

Notably, artificial sweeteners were introduced in consumer products more than four decades ago; however, their overall impact in decreasing sugar consumption, preference and craving has been negligible. This may now be understood at the circuit level (that is, as—in contrast

to sugar—they do not activate the preference circuit), and implies that it may be possible to develop a new class of sweeteners that activate both the sweet-taste receptor in the tongue and the gut–brain axis, and consequently help to moderate the strong drive to consume sugar.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2199-7>.

- Sclafani, A., Marambaud, P. & Ackroff, K. Sucrose-conditioned flavor preferences in sweet ageusic T1r3 and Calhm1 knockout mice. *Physiol. Behav.* **126**, 25–29 (2014).
- de Araujo, I. E. et al. Food reward in the absence of taste receptor signaling. *Neuron* **57**, 930–941 (2008).
- Yarmolinsky, D. A., Zuker, C. S. & Ryba, N. J. P. Common sense about taste: from mammals to insects. *Cell* **139**, 234–244 (2009).
- Zuker, C. S. Food for the brain. *Cell* **161**, 9–11 (2015).
- Elliott, Perry, & Elliott, P. *Production of Sugar in the United States and Foreign Countries* (US Department of Agriculture, 1917).
- Sugar and Sweeteners Yearbook Tables* <https://www.ers.usda.gov/data-products/sugar-and-sweeteners-yearbook-tables/#U.S.%20Consumption%20of%20Caloric%20Sweeteners> (US Department of Agriculture, 2019).
- Nelson, G. et al. Mammalian sweet taste receptors. *Cell* **106**, 381–390 (2001).
- Spector, A. C. & Travers, S. P. The representation of taste quality in the mammalian nervous system. *Behav. Cogn. Neurosci. Rev.* **4**, 143–191 (2005).
- Wang, L. et al. The coding of valence and identity in the mammalian taste system. *Nature* **558**, 127–131 (2018).
- Scott, K. Taste recognition: food for thought. *Neuron* **48**, 455–464 (2005).
- Peng, Y. et al. Sweet and bitter taste in the brain of awake behaving animals. *Nature* **527**, 512–515 (2015).
- Wang, Z., Singhvi, A., Kong, P. & Scott, K. Taste representations in the *Drosophila* brain. *Cell* **117**, 981–991 (2004).
- Sclafani, A., Zukerman, S. & Ackroff, K. Postoral glucose sensing, not caloric content, determines sugar reward in C57BL/6J mice. *Chem. Senses* **40**, 245–258 (2015).
- Ren, X. et al. Nutrient selection in the absence of taste receptor signaling. *J. Neurosci.* **30**, 8012–8023 (2010).
- Zhang, Y. et al. Coding of sweet, bitter, and umami tastes: different receptor cells sharing similar signaling pathways. *Cell* **112**, 293–301 (2003).
- Pérez, C. A. et al. A transient receptor potential channel expressed in taste receptor cells. *Nat. Neurosci.* **5**, 1169–1176 (2002).
- Sclafani, A. Gut–brain nutrient signaling. Appetition vs. satiation. *Appetite* **71**, 454–458 (2013).
- Zukerman, S., Ackroff, K. & Sclafani, A. Post-oral appetite stimulation by sugars and nonmetabolizable sugar analogs. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **305**, R840–R853 (2013).
- Wright, E. M., Loo, D. D. F. & Hirayama, B. A. Biology of human sodium glucose transporters. *Physiol. Rev.* **91**, 733–794 (2011).
- Sheng, M. & Greenberg, M. E. The regulation and function of c-fos and other immediate early genes in the nervous system. *Neuron* **4**, 477–485 (1990).
- Berthoud, H.-R. & Neuhuber, W. L. Functional and chemical anatomy of the afferent vagal system. *Auton. Neurosci.* **85**, 1–17 (2000).
- Sclafani, A. & Glendinning, J. I. Flavor preferences conditioned in C57BL/6 mice by intragastric carbohydrate self-infusion. *Physiol. Behav.* **79**, 783–788 (2003).
- Zukerman, S., Ackroff, K. & Sclafani, A. Rapid post-oral stimulation of intake and flavor conditioning by glucose and fat in the mouse. *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **301**, R1635–R1647 (2011).
- Williams, E. K. et al. Sensory neurons that detect stretch and nutrients in the digestive system. *Cell* **166**, 209–221 (2016).
- Han, W. et al. A neural circuit for gut-induced reward. *Cell* **175**, 665–678.e23 (2018).
- Kaelberer, M. M. et al. A gut–brain neural circuit for nutrient sensory transduction. *Science* **361**, eaat5236 (2018).
- Gunaydin, L. A. et al. Natural neural projection dynamics underlying social behavior. *Cell* **157**, 1535–1551 (2014).
- Guenther, C. J., Miyamichi, K., Yang, H. H., Heller, H. C. & Luo, L. Permanent genetic access to transiently active neurons via TRAP: targeted recombination in active populations. *Neuron* **78**, 773–784 (2013).
- Allen, W. E. et al. Thirst-associated preoptic neurons encode an aversive motivational drive. *Science* **357**, 1149–1155 (2017).
- Callaway, E. M. & Luo, L. Monosynaptic circuit tracing with glycoprotein-deleted rabies viruses. *J. Neurosci.* **35**, 8979–8985 (2015).
- Reardon, T. R. et al. Rabies virus CVS-N2c^{ΔG} strain enhances retrograde synaptic transfer and neuronal viability. *Neuron* **89**, 711–724 (2016).
- Yamamoto, M. et al. Reversible suppression of glutamatergic neurotransmission of cerebellar granule cells in vivo by genetically manipulated expression of tetanus neurotoxin light chain. *J. Neurosci.* **23**, 6759–6767 (2003).
- Chen, T. W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
- Chang, R. B., Strohlic, D. E., Williams, E. K., Umans, B. D. & Liberles, S. D. Vagal sensory neuron subtypes that differentially control breathing. *Cell* **161**, 622–633 (2015).
- Lee, H., Macpherson, L. J., Parada, C. A., Zuker, C. S. & Ryba, N. J. P. Rewiring the taste system. *Nature* **548**, 330–333 (2017).
- Barretto, R. P. J. et al. The neural representation of taste quality at the periphery. *Nature* **517**, 373–376 (2015).
- Mei, N. Vagal glucoreceptors in the small intestine of the cat. *J. Physiol.* **282**, 485–506 (1978).
- Conte, W. L., Kamishina, H. & Reep, R. L. The efficacy of the fluorescent conjugates of cholera toxin subunit B for multiple retrograde tract tracing in the central nervous system. *Brain Struct. Funct.* **213**, 367–373 (2009).
- Dyer, J., Salmon, K. S. H., Zibrik, L. & Shirazi-Beechey, S. P. Expression of sweet taste receptors of the T1R family in the intestinal tract and enteroendocrine cells. *Biochem. Soc. Trans.* **33**, 302–305 (2005).
- Geillinger, K. E. et al. The role of SGLT1 and GLUT2 in intestinal glucose transport and sensing. *PLoS ONE* **9**, e89977 (2014).
- Latorre, R., Sternini, C., De Giorgio, R. & Greenwood-Van Meerveld, B. Enteroendocrine cells: a review of their role in brain–gut communication. *Neurogastroenterol. Motil.* **28**, 620–630 (2016).
- Chambers, A. P., Sandoval, D. A. & Seeley, R. J. Integration of satiety signals by the central nervous system. *Curr. Biol.* **23**, R379–R388 (2013).
- Lein, E. S. et al. Genome-wide atlas of gene expression in the adult mouse brain. *Nature* **445**, 168–176 (2007).
- Armbruster, B. N., Li, X., Pausch, M. H., Herlitze, S. & Roth, B. L. Evolving the lock to fit the key to create a family of G protein-coupled receptors potentially activated by an inert ligand. *Proc. Natl Acad. Sci. USA* **104**, 5163–5168 (2007).
- Gomez, J. L. et al. Chemogenetics revealed: DREADD occupancy and activation via converted clozapine. *Science* **357**, 503–507 (2017).
- Vong, L. et al. Leptin action on GABAergic neurons prevents obesity and reduces inhibitory tone to POMC neurons. *Neuron* **71**, 142–154 (2011).
- Peron, S., Chen, T.-W. & Svoboda, K. Comprehensive imaging of cortical networks. *Curr. Opin. Neurobiol.* **32**, 115–123 (2015).
- Daigle, T. L. et al. A suite of transgenic driver and reporter mouse lines with enhanced brain-cell-type targeting and functionality. *Cell* **174**, 465–480.e22 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Mice

All procedures were carried out in accordance with the US National Institutes of Health (NIH) guidelines for the care and use of laboratory animals, and were approved by the Institutional Animal Care and Use Committee at Columbia University. Adult animals older than 6 weeks of age and from both genders were used in all experiments. C57BL/6J (JAX 000664), *Arc-creER* (TRAP, JAX 021881), TRAP2 (JAX 030323), *TRPM5* KO (JAX 013068), T1R2/3 KO (generated in house, JAX 013065 and 013066), Ai96 (JAX 028866), *VGlut2*-IRES-Cre (JAX 028863), *Gpr65*-IRES-Cre (JAX 029282), *Penk*-IRES2-Cre (JAX 025112), Ai75D (JAX 025106) and R26-TetNT (MGI 3839913).

Fos stimulation and immunohistochemistry

Animals were water restricted for 23 h, given access to 1 ml of water for 1 h, and then water restricted again for another 23 h. The stimulus consisted of 600 mM glucose, 600 mM MDG, 600 mM sucrose, 600 mM 3-OMG, 600 mM galactose, 30 mM AceK or milliQ water for a period of 90 min in the absence of food. For intra-gastric infusion experiments, food was removed from the cage 12 h before stimulus infusion. A syringe pump microcontroller (Harvard Apparatus) was used to deliver 1.5 ml of the stimulus solution at $0.075 \text{ ml min}^{-1}$. After 90 min, mice were perfused transcardially with PBS followed by 4% paraformaldehyde. Brains were dissected, and fixed overnight in paraformaldehyde at 4 °C. The brains were sectioned coronally at 100 μm , and labelled with anti-c-Fos (Santa Cruz, sc-52 goat, 1:500; or SYSY, no. 226004 guinea pig, 1:5,000) in 5% normal donkey serum (EMD Millipore, Jackson ImmunoResearch) in 0.3% Triton X-100 in 1 \times PBS for 48 h at 4 °C with gentle shaking, and then Alexa Fluor 488-, 568- or 647-conjugated donkey anti-goat or anti-guinea pig (Jackson ImmunoResearch) in 5% normal donkey serum in 0.3% Triton X-100 in 1 \times PBS for 24 h at 4 °C with gentle shaking. Images were acquired using an Olympus FluoView 1000 confocal microscope. Larger field-of-view images were acquired using a Nikon AZ100 Multizoom Slide Scanner. Quantification of Fos-labelled neurons was done by manual counting in a $300 \times 300 \mu\text{m}$ region of interest (ROI) in the right cNST.

For intragastric stimulation, animals were anaesthetized with ketamine and xylazine (100 mg kg^{-1} and 10 mg kg^{-1} , intraperitoneal). The stomach was exteriorized through an abdominal incision, and a Silastic (Dow Corning) tubing was inserted into the forestomach region and secured with silk sutures⁴⁹. The other end was tunnelled subcutaneously along the left flank and exteriorized at the dorsal neck area. Mice were individually housed and allowed to recover for at least 5 days before stimulus infusion.

Two-bottle preference assays

The preference-switch experiments were carried out in standard mouse cages holding a custom designed 3D-printed scaffold for two bottles. Each bottle was outfitted with an electronic licking sensor, and access to the licking spout was controlled by a mechanical shutter. Mice were not water deprived before the experiment and had ad libitum access to food throughout. For behavioural tests, mice were first tested for their initial preference by completing 100 drinking trials. Each trial consisted of a choice between 600 mM glucose (or 600 mM MDG) and 30 mM AceK. Trials lasted 5 s and were initiated after the first lick to either bottle; inter-trial intervals were 40 s. To familiarize animals with the two choices, mice were required to complete 500 licks to 600 mM glucose alone, and 30 mM AceK alone; this was repeated twice. Animals were tested for their sugar (or MDG) versus sweetener preference over 36 h using 5-s trials. Preference indexes: Pre, the number of licks to glucose divided by the total number of licks during the first 100 trials of baseline measurements; Post, the number of licks to glucose divided by the total number of licks during the last 100 trials of the behavioural session. Because T1R2/3 double knockouts cannot taste sweet, they

are often averse to the 'bitter' in high concentrations of AceK (that is, not being countered by its high sweetness), therefore they were tested with 300 mM sucrose versus 5 mM AceK.

Molecular cloning of custom pAAV constructs

pAAV.hSyn.FLEX-eGFP-Rpl10a.WPRE.hGH-pA is constructed by ligation of two fragments: the eGFP pAAV backbone fragment was generated by digestion of pAAV-FLEX-EGFP10a, a gift from N. Heintz, A. Nectow and E. Schmidt (Addgene plasmid 98747), with MluI and KpnI, and the hSyn fragment with corresponding restriction ends was generated from pAAV-hSyn-DIO-hM4D(Gi)-mCherry, a gift from B. Roth (Addgene plasmid 44362).

pAAV.CBA.FLEX-GFP-TetX.WPRE.bGH-pA (TetTox) DNA is a gift from P. Wolff⁵⁰.

All pAAVs were amplified in *recA1⁻* NEB Stable cells and extracted by maxiprep (Zymo Research), and serotype 2/9 AAVs were produced by the Janelia Farms viral core.

Genetic access to sugar-preference neurons

The TRAP^{28,29,51} strategy was used in TRAP2 mice to gain genetic access to sugar-activated neurons in the cNST. The 4-hydroxytamoxifen (4OHT, Sigma H6278) was prepared as previously described⁵¹. AAV-injected TRAP2 mice were singly housed, water restricted for 23 h, given access to 1 ml of water for 1 h, water restricted again for another 23 h, and then presented with 600 mM glucose (or 30 mM AceK) ad libitum, in the absence of food and nesting material. After 1 h, mice were injected intraperitoneally with 12.5 mg kg^{-1} 4OHT, and placed back to the same cage for an additional 3 h. At the end of the 4 h of sugar or AceK exposure, mice were returned to regular home-cage conditions (group-caged, with nesting material, ad libitum food and water). Mice were used for experiments a minimum of 7 days after this TRAP protocol.

Stereotaxic surgery

For stereotaxic injections of reporter virus, mice were anaesthetized with ketamine and xylazine (100 mg kg^{-1} and 10 mg kg^{-1} , intraperitoneal), and placed into a stereotaxic frame with a closed-loop heating system to maintain body temperature. For retrograde monosynaptic tracing, animals were unilaterally injected with 100 nl of a 1:1 mixture of AAVs carrying Cre-dependent rabies TVA and glycoprotein G (AAV1 EF1a-FLEX-TVA-mCherry, UNC vector core, and AAV1 FLEX-nGFP-2A-N2c(G) (a gift from T. Reardon)³¹, and a pseudotyped rabies virus carrying dsRed (RABV N2C(Delta G)-dsRed-EnvA, a gift from T. Reardon)³¹. cNST coordinates (Paxinos stereotaxic coordinates⁵²) used for injections are relative to bregma and skull surface: caudal 7.5 mm, lateral $\pm 0.3 \text{ mm}$, ventral 3.7–4 mm.

Monosynaptic retrograde tracing and silencing experiments

For retrograde monosynaptic tracing, Arc-CreER (TRAP)²⁸ mice were allowed to recover 3 weeks after AAV injection, and the TRAP procedure was carried out as described above, except that 4OHT was prepared in corn oil²⁸, and was injected 1 h before stimulus presentation. After 7 days, EnvA-RABV was injected into the same site. Mice were euthanized 2 weeks after the RABV injection and examined for expression of starter cells (nGFP and dsRed) and their monosynaptic inputs (dsRed)^{30,31}.

For synaptic inhibition experiments, sugar-TRAP cNST neurons were bilaterally injected with 300 nl of AAV carrying Cre-dependent TetTox (AAV9 CBA.FLEX-TetTox)⁵⁰.

Synaptic-silencing experiments

C57BL/6J and Trp2^{29,51} mice expressing TetTox in the cNST were tested in the two-bottle sugar versus sweetener preference assay for 48 h. For the first day, mice were acclimatized by exposure to AceK versus water, the second they were given glucose versus water, and the third and fourth days they were tested for their preference to sugar versus sweetener. To ensure silencing did not affect sweet-taste detection,

Article

mice were also examined for their attraction to sugar versus water (second day) as well as artificial sweetener versus water (first day). Fraction consumed for sugar versus AceK on days 3–4 were calculated as (volume of glucose consumed)/(total volume consumed). Fraction consumed for AceK versus water was calculated as (volume of AceK consumed)/(total volume consumed).

Fibre photometry, gut stimulus delivery and vagotomy

Vglut2-cre;Ai96 animals were placed in a stereotaxic frame and implanted with a 400 μm core, 0.48 NA optical fibre (Doric Lenses) 50–100 μm over the right cNST. Photometry experiments were conducted a minimum of 13 days after fibre implantation surgery. Real-time population-level GCaMP fluorescence was recorded using a RZ5P fibre photometry system with Synapse software (Tucker Davis Technologies) as described previously⁵³. In brief, sinusoidally modulated 465 nm and 405 nm light from light-emitting diodes (Doric Lenses) were combined via a multi-port fluorescence mini-cube into a fibre patch-cord connected to the mouse, and real-time demodulated emission signals were saved offline for analysis. Calcium-dependent signals $F_{465\text{nm}}$ were compared with calcium-independent GCaMP fluorescence $F_{405\text{nm}}$ to control for movement and bleaching artefacts. The data was de-trended by first applying a least-squares linear fit to produce $F_{\text{fitted } 405\text{nm}}$, and dF/F was calculated as $(F_{465\text{nm}} - F_{\text{fitted } 405\text{nm}})/F_{\text{fitted } 405\text{nm}}$ (ref.²⁷). Data from each mouse were then normalized to peak fluorescence (calculated as a 10-s window around the peak point), and presented as normalized responses. For each stimulus, the normalized two-trial average was plotted and smoothed over a moving average. To quantify effects of vagotomy, we calculated the ratio of stimulus-related peak amplitude of the normalized trace (within 120 s of stimulus onset) before and after the procedure.

To deliver intestinal stimuli, all animals were anaesthetized with ketamine and xylazine (100 mg kg⁻¹ and 10 mg kg⁻¹, intraperitoneal), re-dosing was performed as necessary with ketamine only (33 mg kg⁻¹). Mice were immobilized as previously described³⁶, positioned in a supine position, with the head rigidly secured using the metal bar. To ensure a clear airway, the mouse was tracheotomized. An incision was made into the greater curvature of the stomach, the tip of the catheter was inserted past the pyloric sphincter and secured by a suture into the duodenal bulb. Another suture was tied around the catheter and stomach to prevent spillage of gastric contents. Upon implantation of the catheter, the intestines were filled with 1 ml of PBS and an exit port cut at the most distally inflated intestinal segment, approximately 12 cm from the catheter. The intestines were flushed with PBS for 5 min at 150 $\mu\text{l min}^{-1}$ before the beginning of each experiment. Stimulus delivery was performed via a series of peristaltic pumps (BioChem Fluidics) operated via custom Matlab software/Arduino microcontroller. Stimuli and washes were delivered through separate lines that converged on a common perfusion manifold (Warner Instruments) connected to the duodenal catheter. All trials were 7-min long and consisted of a 120-s baseline (PBS 150 $\mu\text{l min}^{-1}$), a 60-s stimulus (200 $\mu\text{l min}^{-1}$), and a 4-min washout period (180 s at 600 $\mu\text{l min}^{-1}$, and 60 s at 150 $\mu\text{l min}^{-1}$). Stimuli were each presented twice in an interleaved fashion. All chemicals were obtained from Sigma and dissolved in 1 \times PBS at the following concentrations: 30 mM AceK, 500 mM glucose and 500 mM MDG.

The vagotomy procedure was carried out immediately after the first round of stimuli. Salivary glands were cauterized and removed. Then, skin around the tracheotomy tube was retracted to expose the cervical trunk of the vagus nerve running in close proximity to the carotid artery. The nerve was carefully dissected from the underlying vessels using fine Dumont forceps and fully transected by a pair of Vannas scissors⁵⁴.

Genetic vagal silencing experiments

Vglut2-cre animals were anaesthetized with ketamine and xylazine (100 mg kg⁻¹ and 10 mg kg⁻¹, intraperitoneal). Ophthalmic ointment was applied to the eyes, and subcutaneous injections of carprofen

(5 mg kg⁻¹) and buprenorphine (0.05 mg kg⁻¹) were given to each mouse before surgery. The skin under neck was shaved and betadine and alcohol were used to scrub the skin three times. A midline incision (~1.5 cm) was made and the trachea and surrounding muscles were gently retracted to expose the nodose ganglia. Then, AAV9 CBA.FLEX-TetTox (600 nl per ganglion) containing Fast Green (Sigma, F7252-5G) was injected in both left and right ganglia using a 30° bevelled glass pipette and Nanolitre 2000 microinjector positioned with a micromanipulator. Virus was injected in 60-nl pulses and ganglion targeting was visualized with the dye. At the end of surgery, the skin incision was closed using 5-0 absorbable sutures (CP medical, 421A). Mice were allowed to recover for a minimum of 26 days before behavioural testing. We note that 50% of the animals survived the surgical procedure and bilateral injections.

Vagal calcium imaging

Vglut2-cre;Ai96 or *Gpr65-Cre*;Ai96 mice were anaesthetized, tracheotomized, and positioned on a surgical platform (Thorlabs breadboard). The nodose ganglion was then exposed by severing the posterior tendon of the digastric muscle, cauterizing the occipital branch of the carotid artery and dissecting the trunk of the nerve. Then the preparation was affixed to a set of manual goniometric stages (Newport Instruments) allowing for angular rotation about the longitudinal and lateral axes for optimal positioning under the microscope. Imaging was as previously described³⁵. Imaging data was obtained using an Evolve 512 EMCCD camera (Photometrics). Data was acquired at 10 Hz. A single field of view was chosen for recording and analysis from each mouse, each containing 80–150 segmented single neurons.

To deliver intestinal stimuli for nodose calcium imaging, the duodenum was also exposed and catheterized as described above. A typical trial was 5 min long and consisted of a 60-s baseline (PBS 150 $\mu\text{l min}^{-1}$), a 10-s (or 60-s) stimulus (200 $\mu\text{l min}^{-1}$), and a 3-min washout period (120 s at 600 $\mu\text{l min}^{-1}$, 30 s at 1,800 $\mu\text{l min}^{-1}$, and 30 s at 150 $\mu\text{l min}^{-1}$). Chemicals were dissolved in PBS: AceK, 30 mM; glucose, 500 mM, MDG, 500 mM; mannitol, 500 mM; galactose, 500 mM; 3-OMG, 500 mM. For SGLT1 blocker experiments, 8 mM phlorizin (Sigma) was dissolved in PBS with 3% w/v 1 M NaOH (0.03 M NaOH final)¹⁸, which was titrated back to pH 7.4. The blocker was used within 30 min of preparation. The intestines were pre-washed with PBS + phlorizin flowing at 150 $\mu\text{l min}^{-1}$ for 5 min before commencing the experiments, glucose 500 mM was diluted in PBS + 8 mM phlorizin.

For retrograde labelling of vagal neurons from the duodenum, recombinant Alexa Fluor 594-conjugated CTB (Invitrogen C34777) was injected into the wall of the intestines. A total of 3 μl of 10 mg ml⁻¹ (1%) CTB was injected across 10 sites in the duodenum (within 2 cm of the pylorus) using a 30° bevelled glass pipette connected to a Nanolitre 2010 microinjector (WPI). The pipette was inserted into the outer muscular layer of the intestines (that is, not lumenally) at an acute angle. Mice were used for calcium imaging experiments 3–5 days after CTB injection.

Calcium-imaging data collection and analysis

Calcium-imaging data collected at 10 Hz was downsampled by a factor of 3, and the images stabilized using the NoRMCorre algorithm⁵⁵. Motion corrected movies were then manually segmented in ImageJ using the Cell Magic Wand plugin (<https://www.maxplanckflorida.org/fitzpatricklab/software/cellMagicWand/>). Only ROIs whose average fluorescence was greater than the surrounding neuropil in more than 10% of frames were used for further analysis. Neuropil fluorescence was subtracted from each ROI with the FISSA toolbox⁵⁶, and neural activity was denoised using the OASIS deconvolution algorithm⁵⁷.

Neuronal activity was analysed for significant stimulus-evoked responses, as described in ref.³⁶, with the following modifications. To determine the baseline to calculate z-scores, traces were smoothed over a 15-s moving window, and a baseline distribution of deviations from the median for each cell over the entire experiment was calculated using

periods preceding the stimulus onset. This baseline was then used to calculate a modified z-score by subtracting the median and dividing by the median absolute deviation. Trials with an average modified z-score above 1.6 for the 90 s following presentation of the stimulus were classified as responding trials, and a cell was required to respond in more than 60% of stimulus trials to be classified as a responder. This criterion was validated against visual identification of responses by independent investigators and accurately identified >90% of the same cells with less than 5% false-positive rate. Only cells that reached at least 2% dF/F for the first two trials of glucose were included in heat maps. Heat maps for each experiment were normalized across stimuli, so different stimuli are directly comparable. We note that there were no significant numbers of MDG-only responses (~95% of the neurons that responded to MDG also responded to glucose; a total of 168 MDG responders were analysed and 159 showed responses to both). For the blocker and control data (Fig. 4, Extended Data Fig. 10) responses were filtered to ensure reliable trials preceding blocker addition (that is, the two responses before blocker addition had to be within 70% of each other).

Chemogenetic-activation experiments

For gain of preference experiments, *Penk-cre* animals were stereotactically injected with 300 nl of AAV carrying Cre-dependent activator DREADD (1–2 × 10¹³ GC ml⁻¹; AAV9 Syn-DIO-hM3Dq-mCherry, Addgene 44361), bilaterally in the cNST. At least 8 days was allowed for recovery and viral expression before behavioural testing. We note that in control studies, we validated that injections into the cNST did not infect vagal neurons. We examined six different ganglia with thousands of neurons and detected a total of only four labelled neurons (see also Fig. 3c for an example with AAV1).

C57BL/6J and *Penk* mice expressing hM3Dq in the cNST were tested in a two-bottle grape versus cherry flavour-preference assay. Grape solution was 0.39 g l⁻¹ Kool-Aid Unsweetened Grape (00043000955635) in 1 mM AceK in milliQ water; cherry solution was 0.9 g l⁻¹ Kool-Aid Unsweetened Cherry (00043000955628) in 2 mM AceK in milliQ water. For the first 48 h, animals were tested for their initial preference (Pre) between solutions. Mice were then exposed to cherry only for 2 × 24 h sessions (days 3 and 5), and grape plus 0.005 g l⁻¹ clozapine dihydrochloride (Hello Bio, HB6129-50mg) for 2 × 24 h sessions (days 4 and 6) for conditioning¹. Mice were then assayed for their preference after the conditioning sessions on days 7 and 8. Initial preference is calculated as the average of days 1–2 (volume of grape solution consumed)/(total volume consumed), and post-conditioning preference is similarly calculated from days 7–8. To demonstrate that the switch is independent of the nature of the initial less-preferred stimuli, 4/8 C57BL/6J and 4/8 *Penk*-hM3Dq mice were tested with reversed flavour conditions (that is, conditioned to 0.9 g l⁻¹ cherry in 1 mM AceK with clozapine).

Brief-access preference assay

C57BL/6J mice were tested for their immediate taste preferences in a short-access two-bottle preference assay⁷. Singly housed naive mice were acclimatized in new cages with access to two bottles of water overnight. Animals were then water deprived for 1h and presented with 600 mM glucose versus 600 mM MDG for 1h. Preference for glucose was calculated as (volume of glucose consumed)/(total volume consumed).

Insulin and glucose measurements

Plasma insulin and blood glucose measurements were performed as previously described⁵⁸. Male C57BL/6J mice were group-housed in cages with wood chip bedding. Mice were habituated to scruffing and blood draws at least twice before the experiment. On the day of sample collection, animals were subjected to a 5-h fast (food removed and transferred to clean cages) beginning at the lights-on period of the light–dark cycle (07:00). Mice were gavaged with 555 mM glucose or MDG at 2 mg kg⁻¹. Blood (~100 µl) was collected before and 15 min after gavage into a chilled heparin-coated Eppendorf tube. Glucose measurements were

taken on whole blood via hand-held glucometer (OneTouch Verio). For insulin measurements, samples were run on a mouse insulin ELISA kit (Mercodia, 10-1247-01) according to the manufacturer's directions.

Retrograde labelling of vagal neurons from brainstem

C57BL/6J mice were stereotactically injected with 50 nl of red or green fluorescent RetroBeads (Lumafluor) in the cNST or Cuneate nucleus (Paxinos stereotaxic coordinates relative to Bregma and skull surface: caudal 7.5 mm, lateral 0.9 mm, ventral 3.6–3.9 mm). Mice were euthanized 6–7 days after RetroBeads injection. Prior to analysis, the brainstem was sliced coronally to confirm accurate targeting to the cNST and Cuneate. Nodose and dorsal root ganglia (across the cervical, thoracic and lumbar segments)⁵⁹ were examined for fluorescent labelling.

Statistics

No statistical methods were used to predetermine sample size, and investigators were not blinded to group allocation. No method of randomization was used to determine how animals were allocated to experimental groups. Statistical methods used include one-way ANOVA followed by Tukey's HSD post hoc test, two-tailed *t*-test, or the two-sided Mann–Whitney *U*-test, and are indicated for all figures. All analyses were performed in MATLAB. Data are presented as mean ± s.e.m.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data supporting the findings of this study are available from the corresponding author upon request.

Code availability

Custom code is available from the corresponding author.

49. Ueno, A. et al. Mouse intragastric infusion (iG) model. *Nat. Protoc.* **7**, 771–781 (2012).
50. Murray, A. J. et al. Parvalbumin-positive CA1 interneurons are required for spatial working but not for reference memory. *Nat. Neurosci.* **14**, 297–299 (2011).
51. DeNardo, L. A. et al. Temporal evolution of cortical ensembles promoting remote memory retrieval. *Nat. Neurosci.* **22**, 460–469 (2019).
52. Paxinos, G. & Franklin, K. *The Mouse Brain in Stereotaxic Coordinates* 2nd edn (Academic, 2001).
53. Lerner, T. N. et al. Intact-brain analyses reveal distinct information carried by SNc dopamine subcircuits. *Cell* **162**, 635–647 (2015).
54. Cyphert, J. M. in *Mouse Models of Allergic Disease: Methods and Protocols* (ed. Allen, I. C.) 219–227 (Humana, 2013).
55. Pnevmatikakis, E. A. & Giovannucci, A. An online algorithm for piecewise rigid motion correction of calcium imaging data. *J. Neurosci. Methods* **291**, 83–94 (2017).
56. Keemink, S. W. et al. FISSA: a neuropil decontamination toolbox for calcium imaging signals. *Sci. Rep.* **8**, 3493 (2018).
57. Friedrich, J., Zhou, P. & Paninski, L. Fast online deconvolution of calcium imaging data. *PLoS Comput. Biol.* **13**, e1005423 (2017).
58. Moriya, R., Shirakura, T., Ito, J., Mashiko, S. & Seo, T. Activation of sodium-glucose cotransporter 1 ameliorates hyperglycemia by mediating incretin secretion in mice. *Am. J. Physiol. Endocrinol. Metab.* **297**, E1358–E1365 (2009).
59. Sleight, J. N., Weir, G. A. & Schiavo, G. A simple, step-by-step dissection protocol for the rapid isolation of mouse dorsal root ganglia. *BMC Res. Notes* **9**, 82 (2016).
60. Martin, G. *Neuroanatomy Text and Atlas* 4th edn (McGraw Hill, 2003).
61. Luo, L., Callaway, E. M. & Svoboda, K. Genetic dissection of neural circuits. *Neuron* **57**, 634–660 (2008).

Acknowledgements We thank N. Ryba for experimental suggestions and helpful comments; R. Barretto for advice on the calcium-imaging pipeline; L. Luo for the TRAP mice; S. Lieberles for GPR65-Cre mice; P. Wulff for the tetanus toxin construct; A. Skowronski and C. Leduc for their assistance in performing blood glucose and insulin measurements; L. Rickman for expert help with figures; R. Lessard for earlier contributions; members of the Zuker lab for helpful discussions; and E. Sobolik, L. Hsin, Y. Zhang, A. Holguin, A. Conomikes, E. Shaw, B. McTyre and J. Li, who participated in various aspects of this work. Imaging was performed with support from the Zuckerman Institute's Cellular Imaging platform. Research reported in this publication was supported in part by the Russell Berrie Foundation program in the neurobiology of obesity (to C.S.Z. and R. Leibel). A.C.S. was supported by the MSTP program, H.-E.T. was supported by the Agency for Science, Technology and Research (A*STAR) of Singapore, and Y.G. was supported by a predoctoral fellowship from NRSA and the MSTP

Article

program. C.S.Z. is an investigator of the Howard Hughes Medical Institute and a Senior Fellow at Janelia Farm Research Campus. Figures were generated with the help of BioRender.

Author contributions A.C.S. and H.-E.T. designed the study, carried out the experiments and analysed data, H.J. performed retrograde tracing experiments and helped with the TRAP system. M. Vignovich analysed calcium-imaging data, and helped to develop the analysis pipeline. M. Villavicencio and K.S.T. designed and characterized engineered animals and behavioural experiments. Y.G. participated in the initial phases of this study. C.S.Z. designed the study and analysed data. C.S.Z., A.C.S. and H.-E.T. wrote the paper.

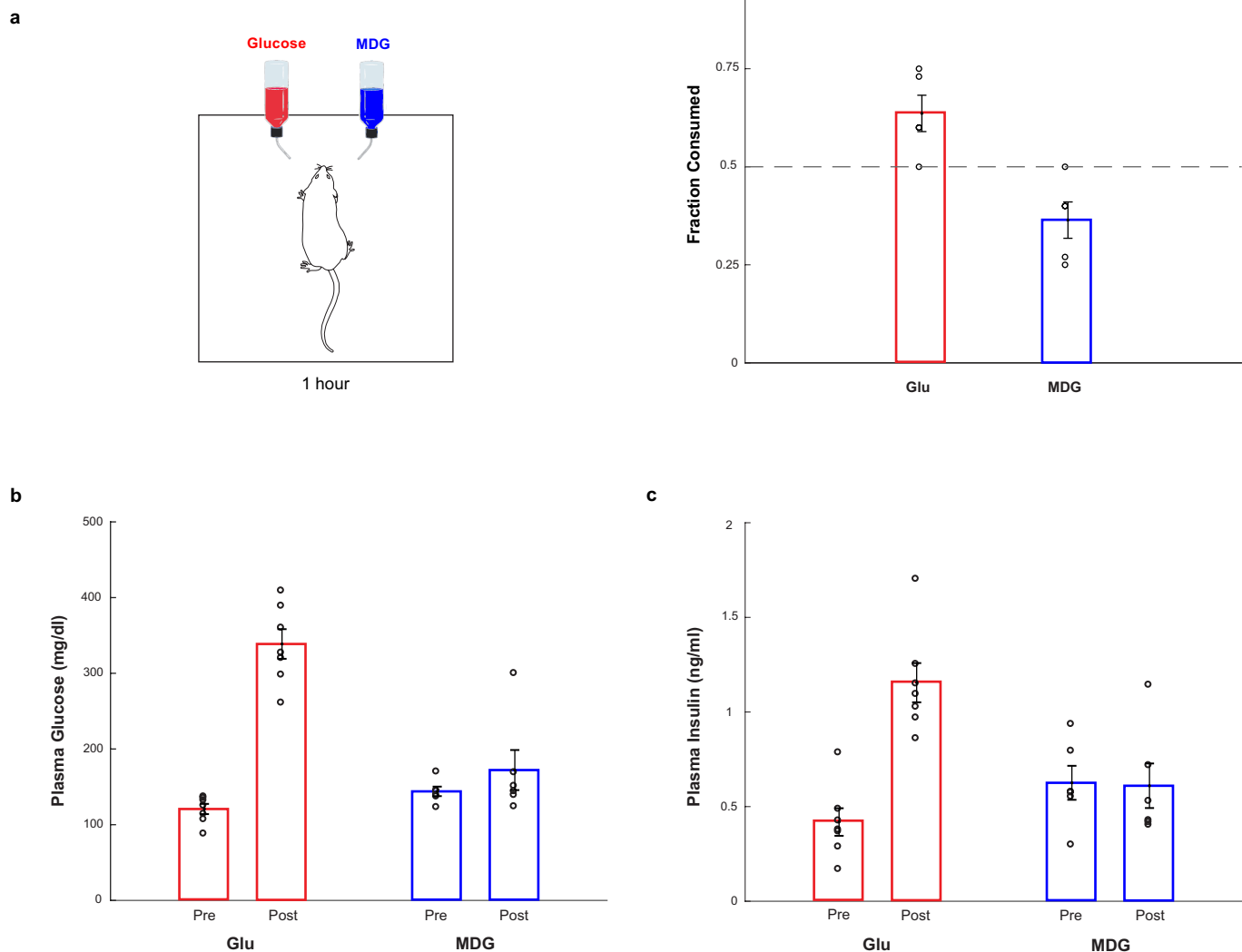
Competing interests C.S.Z. is a scientific co-founder of and advisor to Kallyope. The other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2199-7>.

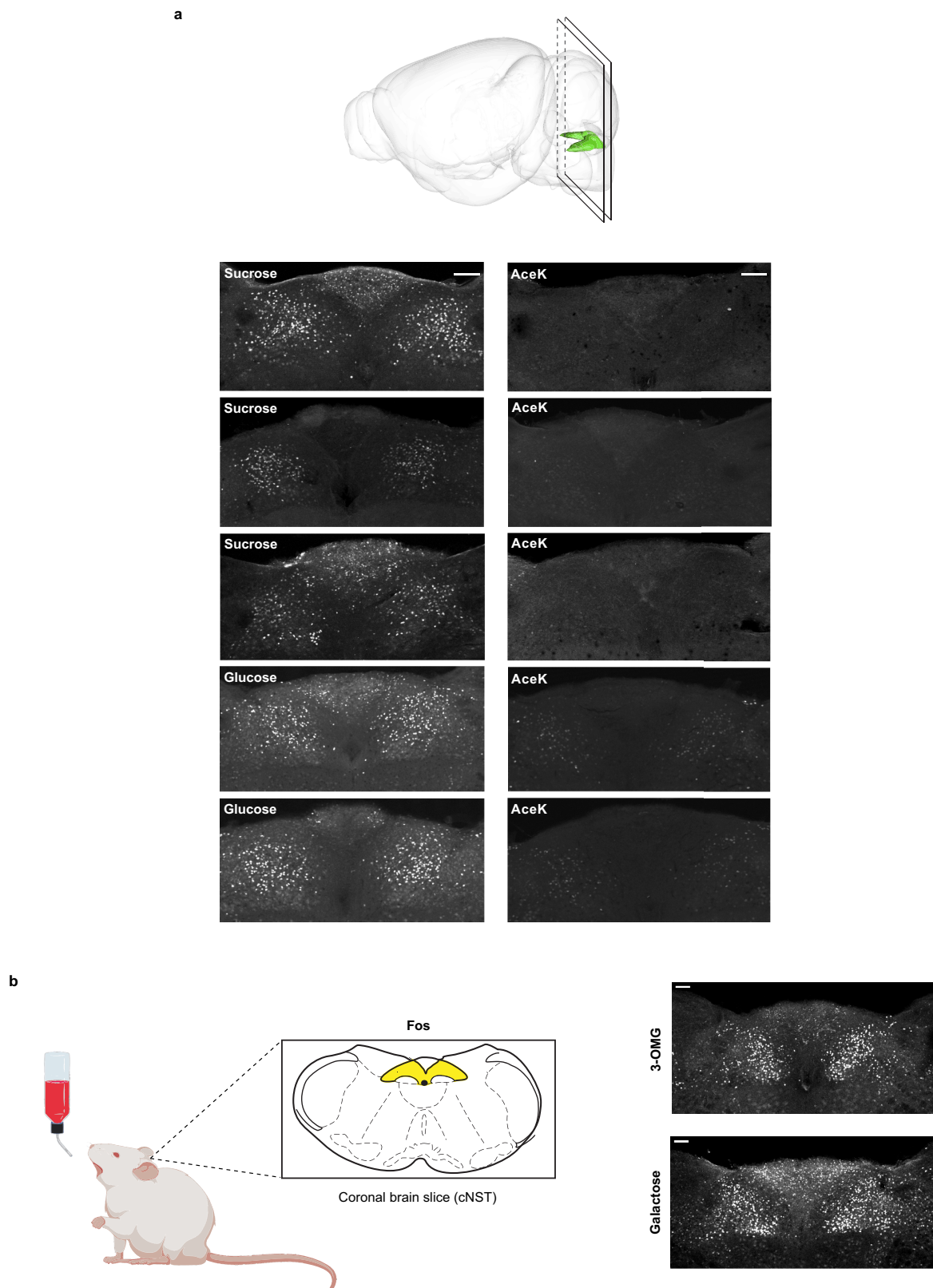
Correspondence and requests for materials should be addressed to C.S.Z.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



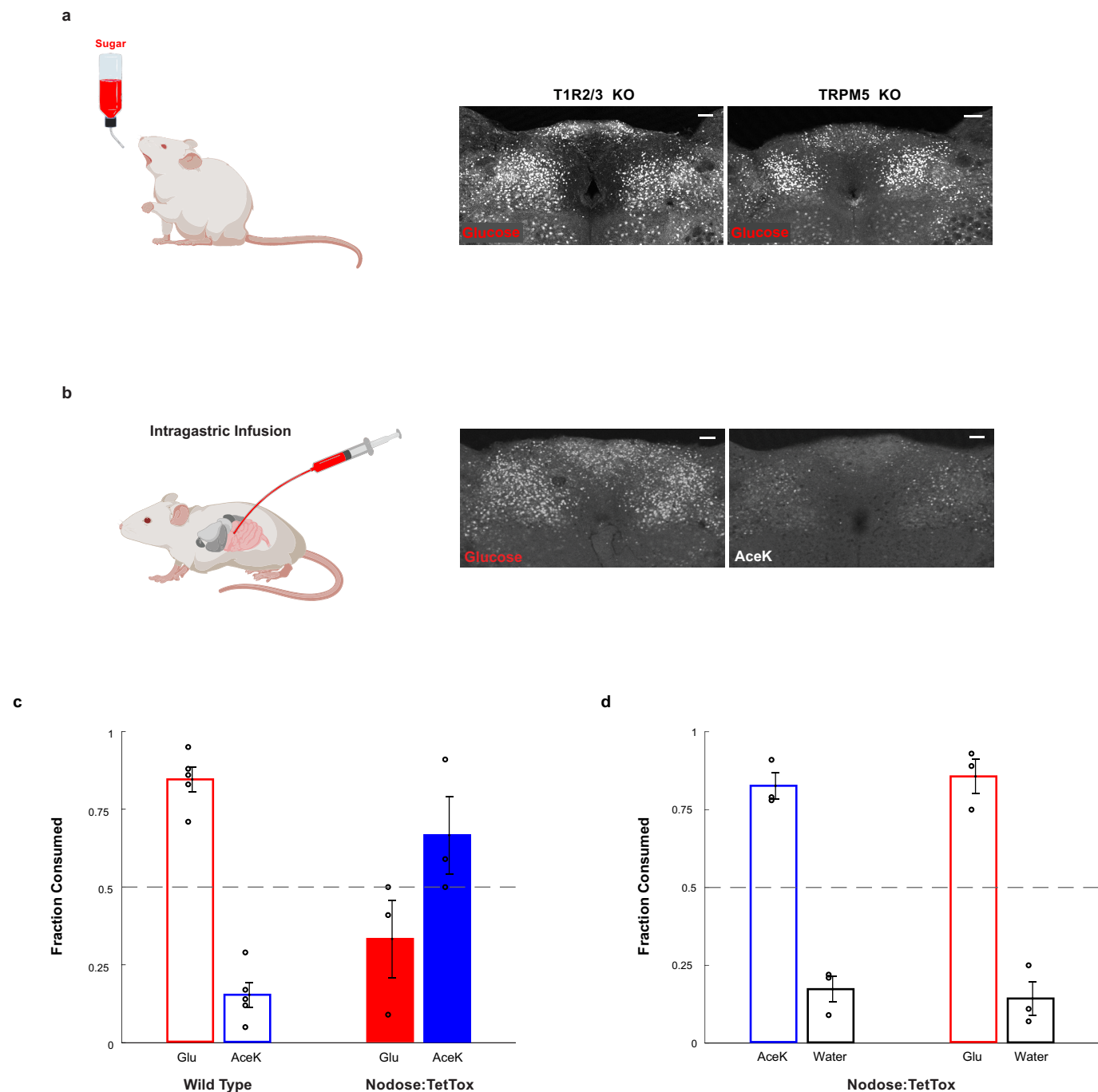
Extended Data Fig. 1 | Glucose and MDG preference. **a**, When mice are given a choice between 600 mM glucose or 600 mM MDG, using a brief-access (1 h) test, naive animals display a small preference for glucose over MDG ($n = 5$, two-tailed paired t -test, $P = 0.0406$), probably because MDG is slightly less sweet and thus not as attractive. Values are mean \pm s.e.m. **b, c**, Although the non-caloric sugar analogue MDG is very effective in causing a preference switch (see Fig. 1), it does not cause increases in plasma glucose or release of

insulin. Mice were gavaged with glucose or MDG, and plasma glucose and insulin levels were sampled before (Pre), and at 15 min after the gavage (Post). **b**, Plasma glucose after glucose gavage (red bars). $n = 7$, two-tailed paired t -test, $P = 4 \times 10^{-5}$. Plasma glucose after MDG gavage (blue bars). $n = 6$, two-tailed paired t -test, $P = 0.36$. **c**, Plasma insulin levels after glucose gavage (red bars). $n = 7$, two-tailed paired t -test, $P = 7 \times 10^{-6}$. Plasma insulin levels after MDG gavage (blue). $n = 6$, two-tailed paired t -test, $P = 0.94$. Values are mean \pm s.e.m.



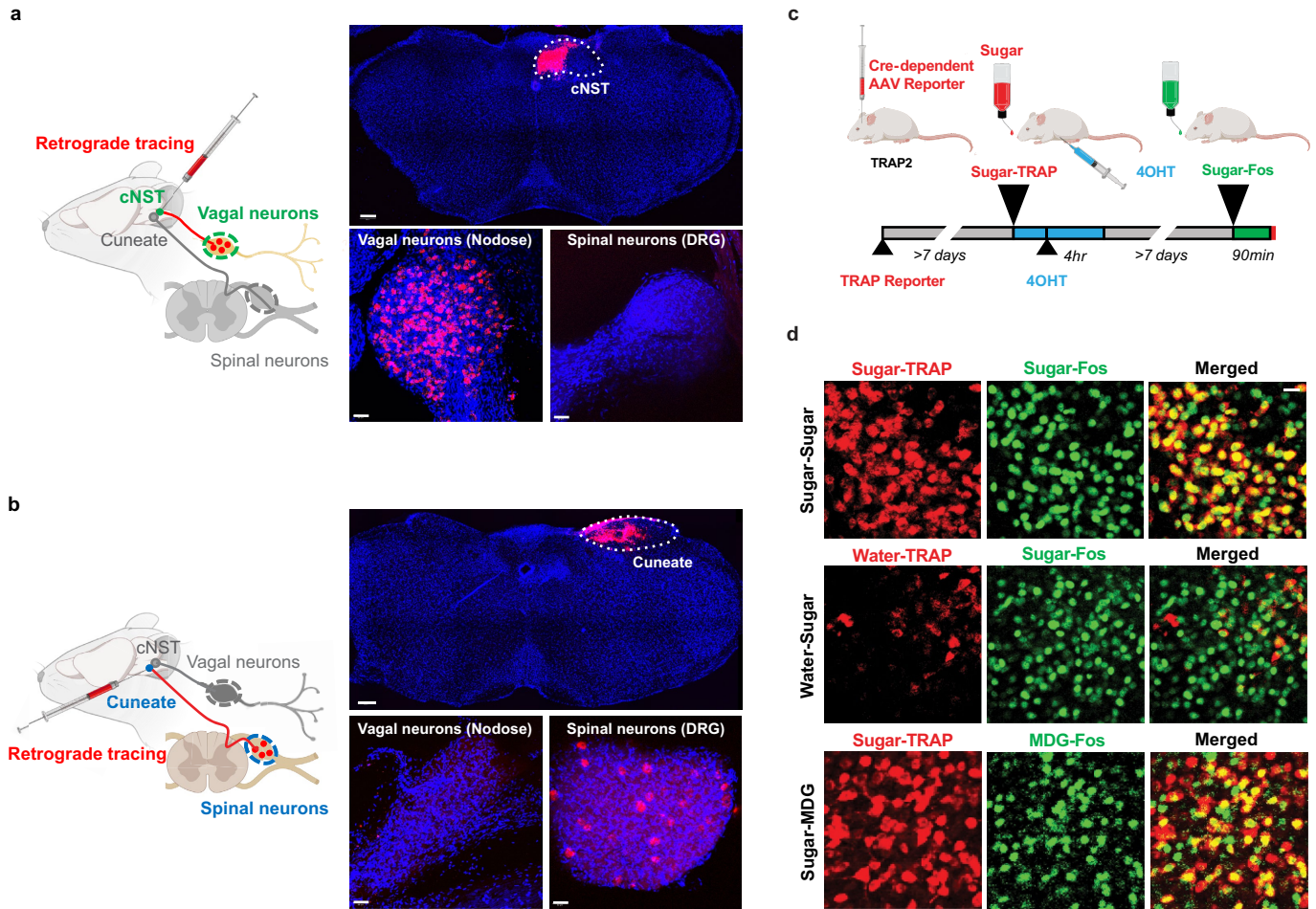
Extended Data Fig. 2 | Fos responses are robust and reliable. a, The brain diagram illustrates the position of the NST and the plane of the sectioning. Shown are cNST sections stained with Fos antibodies after exposing the animals to 90 min of 600 mM sucrose, 600 mM glucose or 30 mM AceK. Each panel is a confocal maximal projection image from Bregma -7.5 mm consisting of 3 sections $15\ \mu\text{m}$ apart. Each panel (sucrose, glucose or AceK)

represents a different animal, $n = 3$ independent experiments. Note the robustness of the signals across animals. See Methods for details. **b,** Mice were stimulated with 600 mM 3-OMG ($n = 6$ mice) or 600 mM galactose ($n = 3$ mice) (see also Fig. 4, Extended Data Fig. 10). Note strong Fos signals in cNST neurons, $n = 2$ independent experiments (total of 9 mice). Scale bars, $100\ \mu\text{m}$.



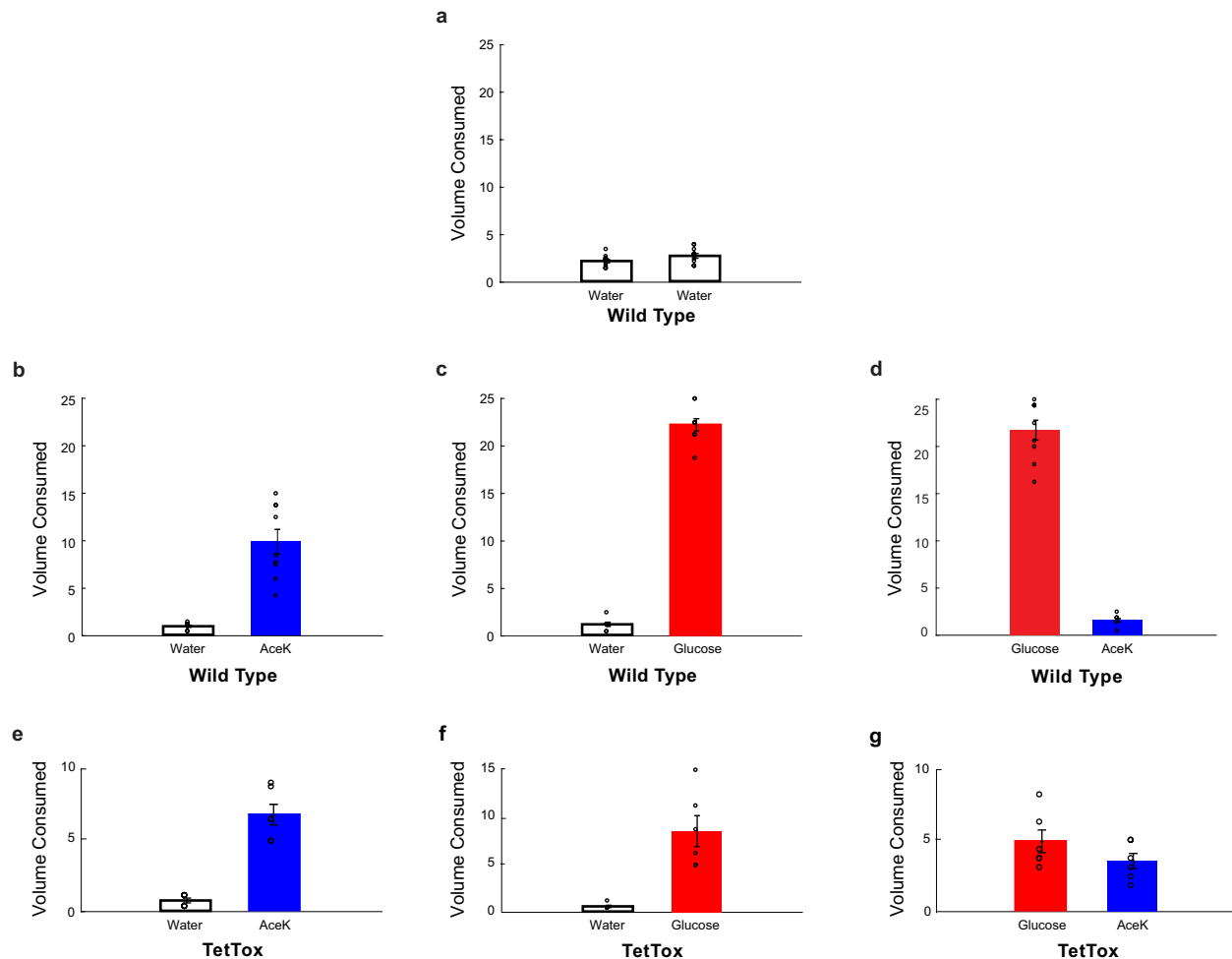
Extended Data Fig. 3 | The development of sugar preference. **a**, Glucose stimulates cNST neurons in mice lacking the sweet taste receptor (T1R2/3^{-/-}), or in mice lacking the TRPM5 ion channel (TRPM5^{-/-}). See Fig. 1e for quantification. T1R2/3^{-/-}, $n = 5$ mice, ANOVA followed by Tukey's HSD post hoc test, $P < 0.0001$; TRPM5^{-/-}, $n = 7$ mice, ANOVA followed by Tukey's HSD post hoc test, $P < 0.0001$. Values are mean \pm s.e.m. Scale bars, 100 μ m. **b**, Direct intragastric infusion of glucose, but not AceK, robustly activates the cNST. $n = 2$ independent experiments. Scale bars, 100 μ m. **c**, **d**, Genetic silencing of vagal sensory neurons. **c**, Sugar-preference graphs for wild-type mice ($n = 5$ mice),

demonstrating the robust development of preference for sugar versus artificial sweetener (see also Fig. 1). By contrast, silencing of the sensory neurons in the nodose ganglia, by bilateral injection of AAV-DIO-TetTox into the nodose ganglia of *Vglut2-cre* mice (see Methods), abolishes the development of sugar preference; $n = 3$ mice, two-sided Mann-Whitney *U*-test, $P = 0.035$. Values are mean \pm s.e.m. **d**, However, silencing vagal sensory neurons does not impair the innate attraction to sweet solutions; shown are behavioural responses to AceK versus water, and glucose versus water ($n = 3$ animals, preference index for AceK = 0.82, preference index for glucose = 0.85). Values are mean \pm s.e.m.



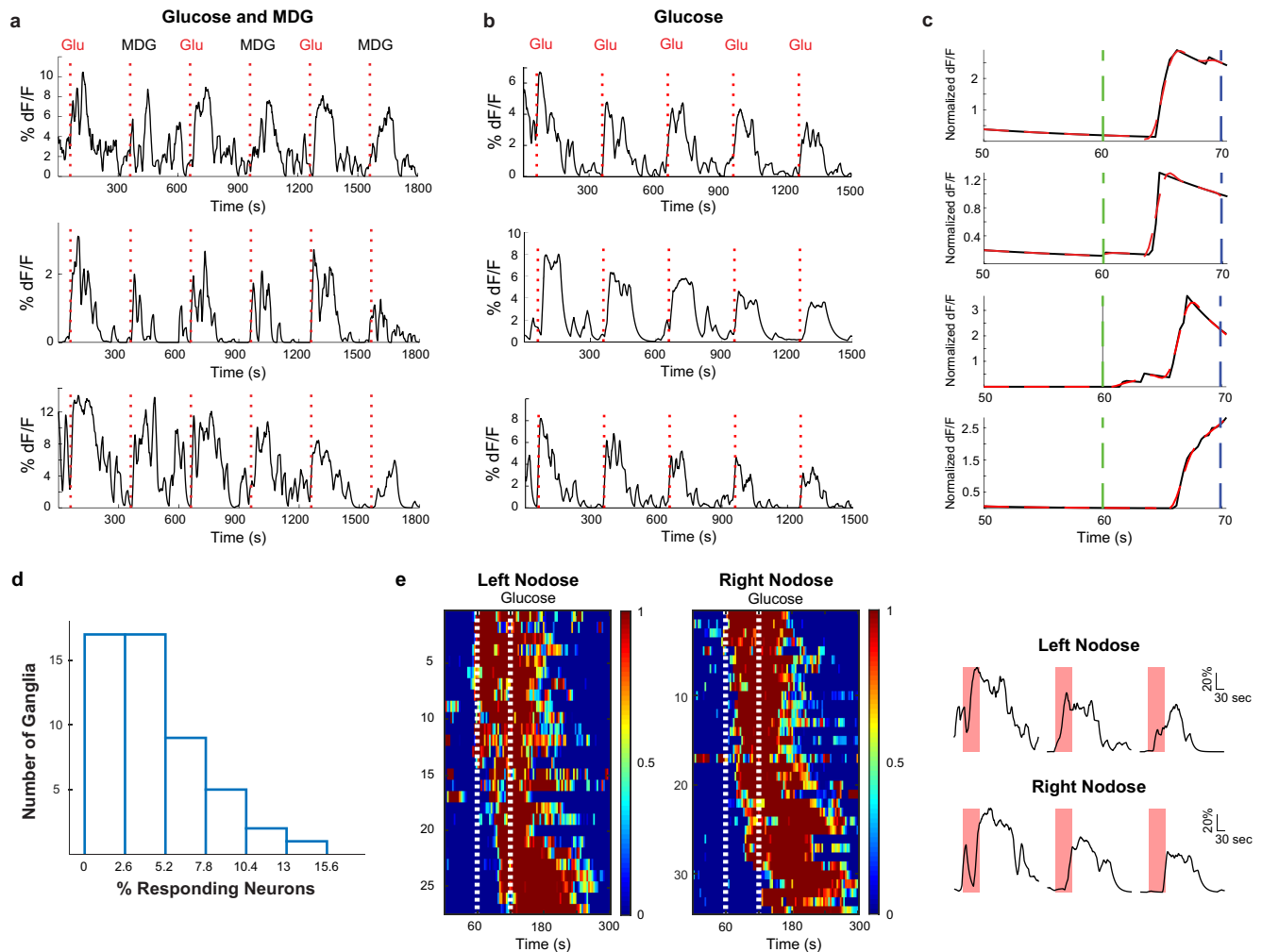
Extended Data Fig. 4 | Retrograde labelling from cNST. **a**, A fluorescent retrograde tracer (red RetroBeads, Lumafluor) was stereotactically injected into the cNST to label its inputs. The nodose ganglia and dorsal root ganglia were checked for transfer of the fluorescent label after 6–7 days. The nodose ganglion (vagal neurons), but not the dorsal root ganglion (spinal neurons), was robustly labelled⁶⁰. $n = 2$ independent experiments. **b**, RetroBeads were also injected into the cuneate nucleus, a brainstem area near but distinct from the cNST. Vagal neurons were not labelled. By contrast, note robust labelling of spinal neurons ($n = 2$ independent experiments). Nuclei were counterstained with DAPI (blue). Scale bars, 200 μm (Brainstem), 50 μm (nodose, DRG). **c**, Validation of TRAPing procedure to confirm that the sugar-activated cNST neurons marked by the expression of Fos are the same as the ones labelled by

Cre recombinase in the genetic TRAPing experiments. We genetically labelled the sugar-induced TRAPed neurons with a Cre-dependent fluorescent reporter⁶¹, and then performed a second cycle of sugar stimulation followed by Fos antibody labelling. **d**, Top, neurons labelled by the Cre-dependent reporter after sugar TRAPing ('sugar-TRAP', pseudocoloured red) are the same as those labelled by Fos after a second cycle of sugar stimulation ('sugar-Fos', green; see Methods and text for details), >80% of Sugar-Fos neurons are also sugar-TRAP positive ($n = 7$ animals). Middle, note that the few neurons labelled after water-TRAP in response to water do not overlap with those labelled with Fos antibodies after sugar stimulation. Bottom, the sugar-TRAP neurons are also activated by the non-caloric sugar analogue MDG; >80% of MDG-Fos are sugar-TRAP positive. Scale bar, 20 μm .



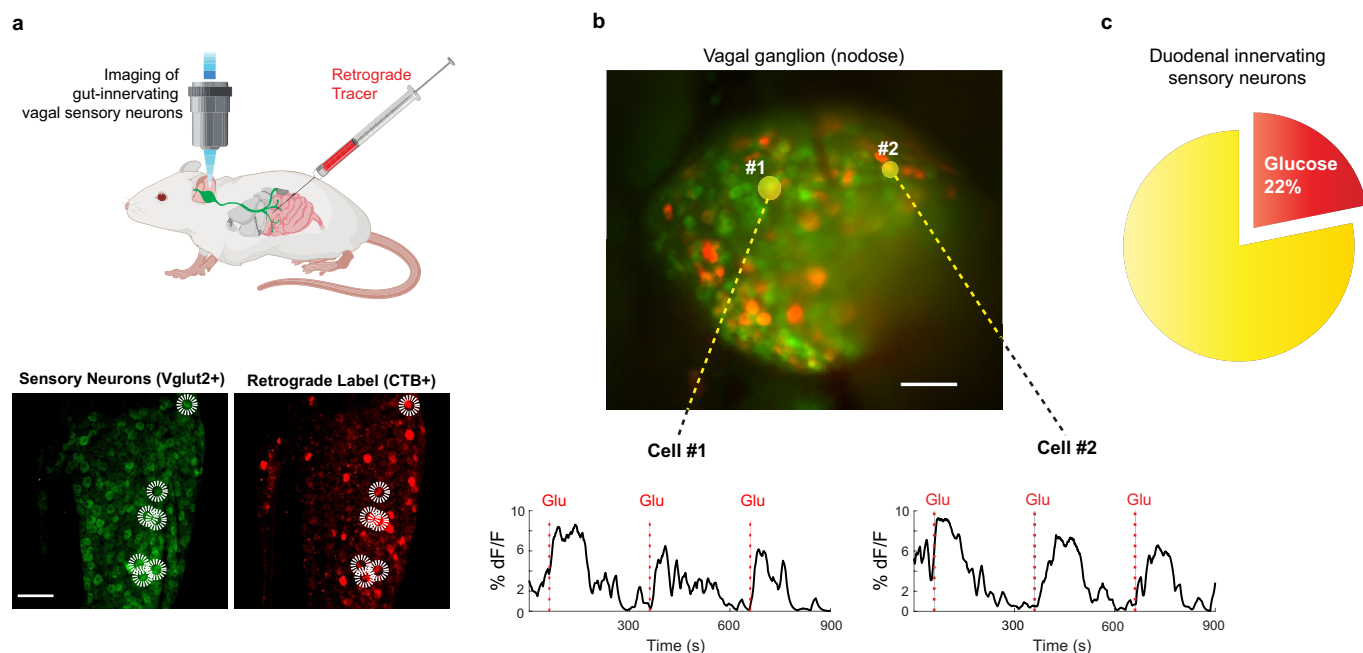
Extended Data Fig. 5 | Mice with a silenced sugar-preference circuit behave as normal mice, drinking artificial sweeteners. a, A normal, non-thirsty mouse drinks about 5 ml of water during a 24-h window. $n=11$ mice. Values are mean \pm s.e.m. **b**, If presented with a sweet option (but not sugar, so as to not create a preference) they show a small but significant increase in total volume consumed, but now most of the total consumption is from the sweet choice rather than water ($n=9$ animals, two-tailed paired t -test, $P=1 \times 10^{-4}$). Values are mean \pm s.e.m. **c**, By contrast, if the options are water versus sugar, so that it creates a preference, they massively increase total volume consumed, and nearly all is from the sugar solution ($n=9$ animals, two-tailed paired t -test,

$P=3 \times 10^{-10}$). Values are mean \pm s.e.m. **d**, As expected, wild-type controls develop a strong preference for sugar versus AceK ($n=9$ animals, two-tailed paired t -test, $P=3 \times 10^{-8}$). Values are mean \pm s.e.m. **e**, **f**, Mice with the preference circuit silenced behave as control animals presented with a sweet, non-preference creating choice (compare **e**, **f** with **b**) ($n=6$ mice, two-tailed paired t -test, $P=6 \times 10^{-4}$ for AceK, $P=4 \times 10^{-3}$ for glucose). Values are mean \pm s.e.m. **g**, Silenced animals consumed nearly equal volumes of sugar and artificial sweetener ($n=6$ animals, two-tailed paired t -test, $P=0.1$). Values are mean \pm s.e.m.



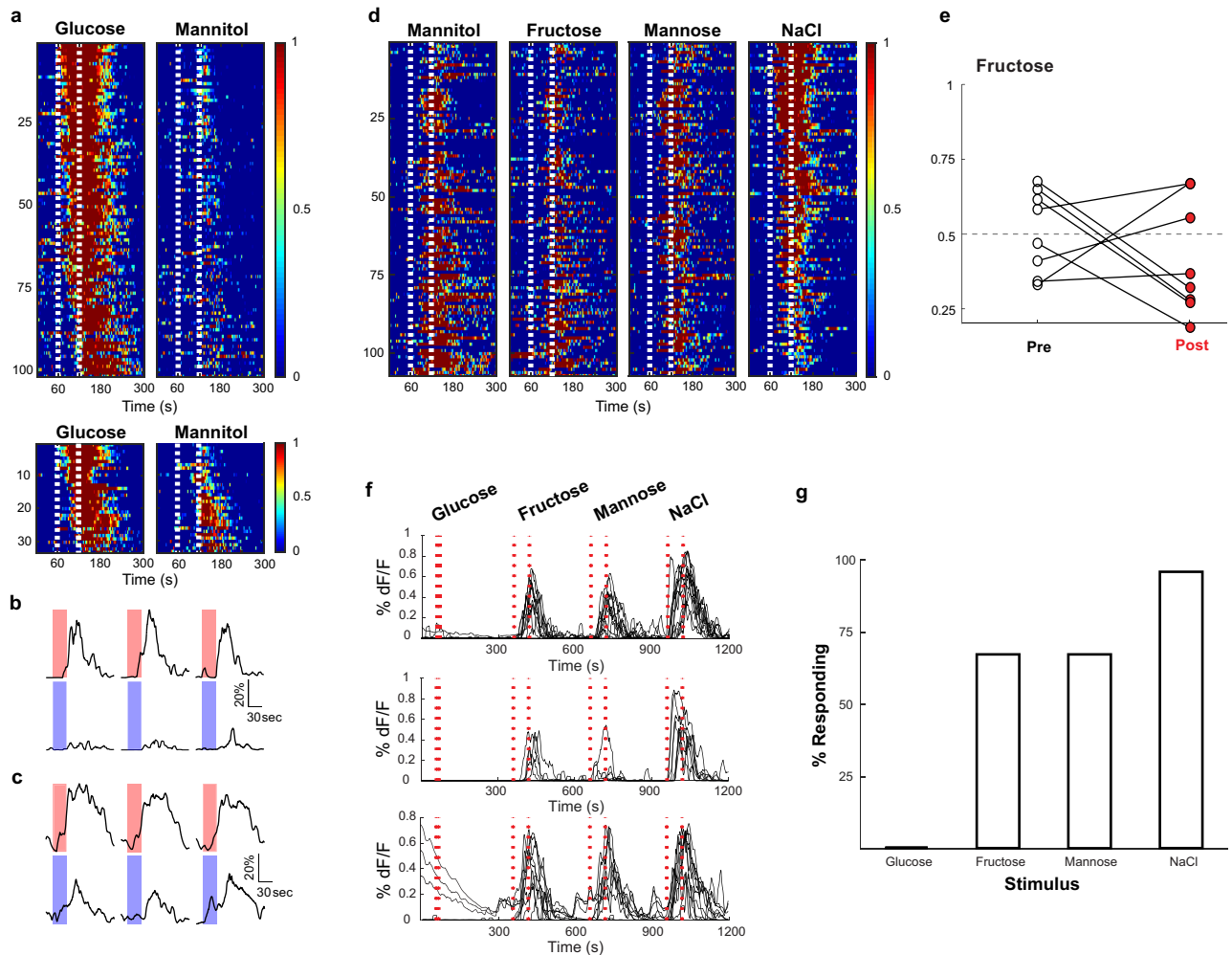
Extended Data Fig. 6 | Vagal-neuron responses to sugar and MDG are highly reproducible and timed-locked to the stimulus. **a**, Shown are vagal-neuron responses to 6 consecutive 10-s intestinal stimuli of alternating trials with 500 mM glucose and 500 mM MDG (stimulus delivery and timings are as described in the Methods). Each of the sample traces depicts the response from a different neuron. **b**, Shown are vagal-neuron responses to 5 consecutive 10-s intestinal stimuli with 500 mM glucose (stimulus delivery and timings are as described in the Methods). Each of the sample traces shows the response from a different neuron. **c**, Expanded time scale of responses to the 10-s 500 mM glucose stimulus from 10 s before to 10 s after termination of the stimulus. The green dashed lines indicate the initiation of the stimulus, and the blue dashed lines denote termination of the 10-s stimulus. Calcium responses are shown in solid black and exponential fits to the response latency and kinetics are shown in red. Note responses time-locked to stimulus delivery; the top two traces depict two cells from two different mice in response to glucose, and the bottom two traces depict two cells from two different mice in response to MDG; latencies varied between 3 and 6 s, and were within the 10-s stimulation

window. Some cells exhibited longer latencies (see for example, heat maps in Fig. 4, Extended Data Fig. 8). We believe the cells with longer response latencies may represent intestinal glucose responders located farther down the intestinal segment, and thus would be expected to demonstrate longer latencies³⁷. **d**, On average, approximately 5% of vagal neurons respond reliably to a 10-s 500 mM glucose stimulus. The histogram shows the percentage of GCaMP-expressing vagal neurons responding to the 10-s glucose stimulus. Average = $4.6 \pm 0.05\%$ ($n = 4,803$ neurons from 51 ganglia, mean \pm s.e.m.). **e**, Recent findings²⁵ have suggested that appetitive behavioural responses are elicited through stimulation of vagal terminals originating from the right nodose ganglion. Shown are heat maps depicting z-score normalized average calcium responses of individual ganglion neurons after a 60-s pulse of 500 mM glucose. We observe no differences in responses to intestinal glucose from either the left or right vagal ganglia. Also shown are example traces from different neurons from the left and right Nodose ganglion; red bars indicate the 60-s stimulus; scale bars indicate percentage maximal response.



Extended Data Fig. 7 | Vagal neurons innervating duodenal segment sense sugar. **a**, Top, schematic of retrograde tracing experiment. Fluorescently conjugated CTB³⁸ was injected into the proximal duodenum to back fill and label the cell bodies of duodenum-projecting vagal neurons (z-projection of $n = 22$ confocal planes from a representative ganglion, see Methods for details). The two bottom panels show a sample retrogradely labelled ganglion with sensory neurons (*Vglut2-cre* driving the GCaMP reporter) marked in green (left) and those labelled by CTB marked by red fluorescence (right). Double-positive neurons are highlighted by the white circles. Scale bar, 100 μm . **b**,

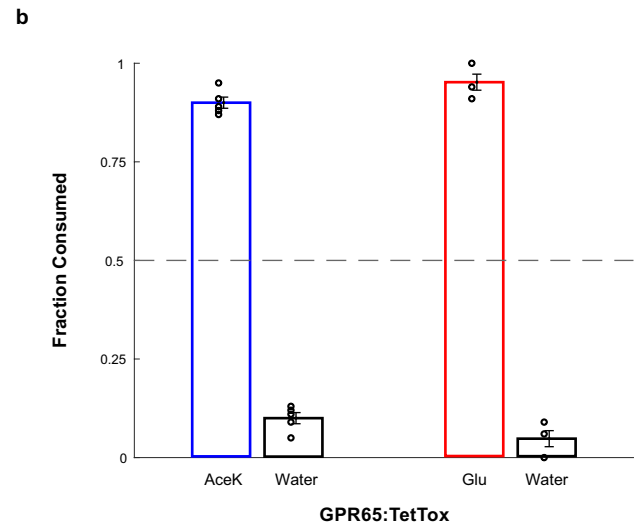
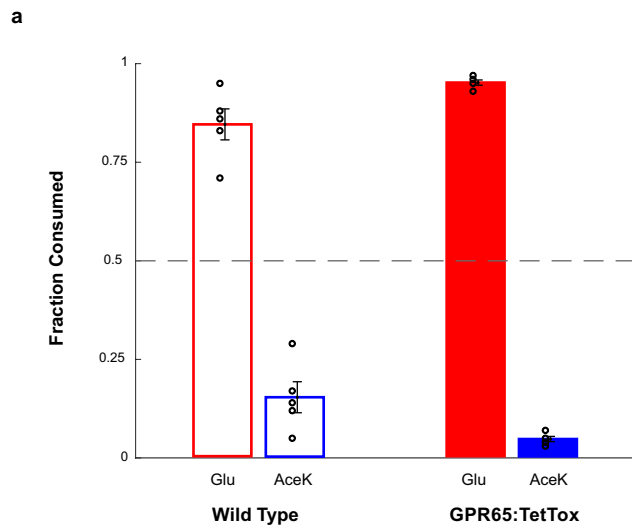
Representative field of a vagal imaging session showing the overlay of CTB and GCaMP. The two yellow circled neurons (denoted as #1 and #2) were labelled by retrogradely applied CTB in the duodenal segment, and exhibited strong responses to glucose ($n = 16$ ganglia from 10 mice). Scale bar, 100 μm . **c**, A total of 12 out of 55 double-positive neurons responded to the 10-s glucose stimulus (see Extended Data Fig. 6d for a comparison with uninjected animals). $n = 16$ ganglia from 10 mice. Note the substantial enrichment in the number of responders when pre-tagged by retrograde labelling: ~20% in the duodenal tagged versus 4–5% in the whole population.



Extended Data Fig. 8 | Glucose responders are not sensing osmolarity.

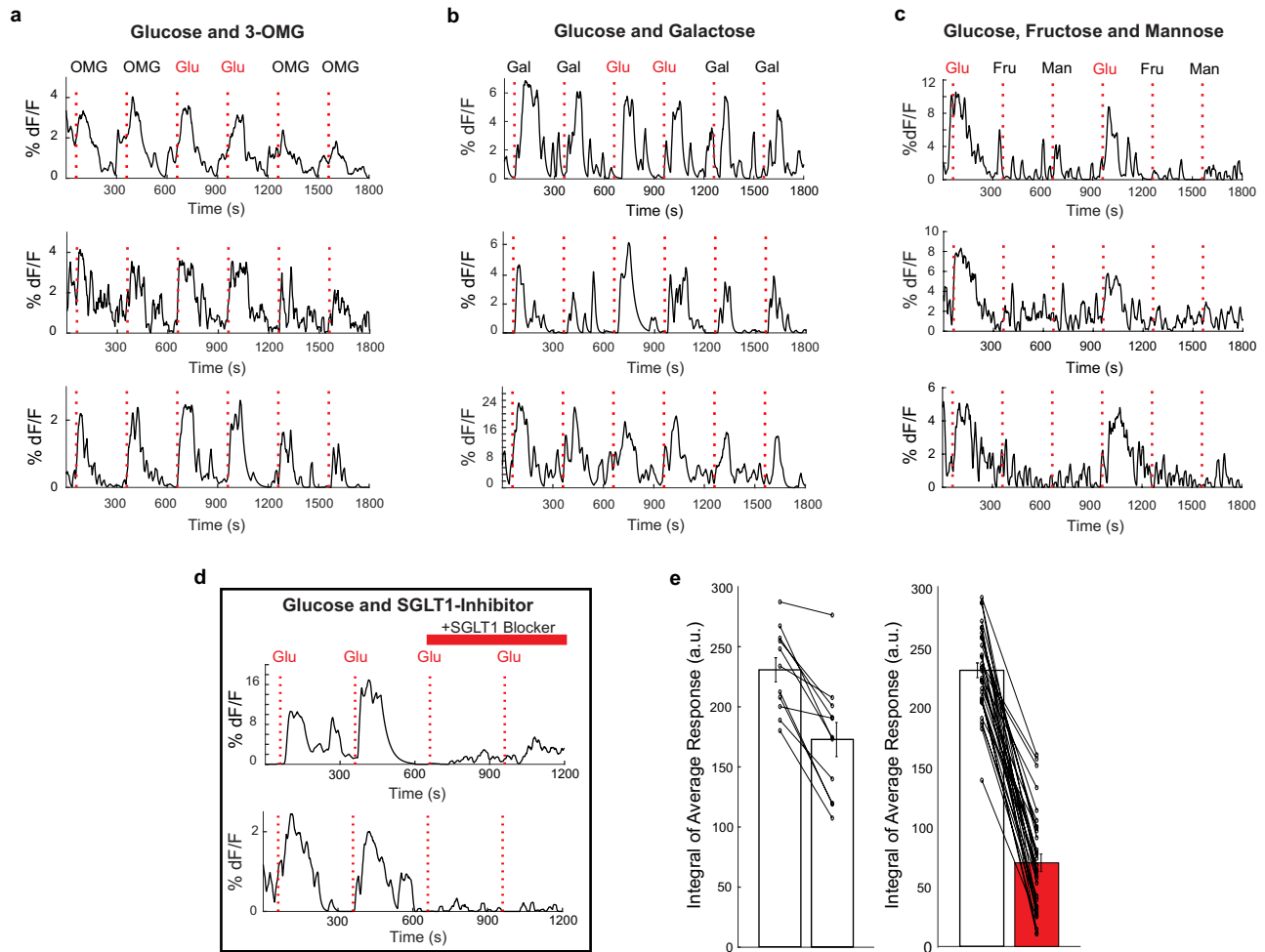
Williams et al.²⁴ identified vagal neurons that indiscriminately responded to high concentrations of several stimuli delivered in very large stimulus volume for hundreds of seconds. We believe these responses, largely independent of the quality of the stimulus, are intestinal osmolarity signals. **a**, Shown are heat maps summarizing responses to interleaved 60-s stimuli of 500 mM glucose and 500 mM mannitol. Each row represents the average activity of a single cell during three interspersed exposures to the stimulus. Stimulus window is indicated by the dashed white lines. Of 134 neurons that responded to intestinal application of 500 mM glucose for 60 s, 101 did not exhibit statistically significant responses to mannitol (top). However, 33 (~25%) showed responses to both 500 mM glucose and 500 mM mannitol (bottom). $n = 5$ mice. When the intestinal stimulus consisted of a short pulse (that is, 10 s; 33 μ l volume) no responses were detected to 500 mM mannitol (data not shown). **b**, Sample traces (three trials each) of a neuron responding to glucose (red) but not mannitol (blue). **c**, Sample traces (three trials each) of a neuron responding to glucose and mannitol. Scale bars indicate percentage maximal response.

d, Heat maps showing responses to a 60-s stimuli of 1 M mannitol, 1 M fructose, 1 M mannose and 1 M NaCl. Note that the same cells respond indiscriminately to the various stimulus ($n = 4$ mice). **e**, The graph shows preference plots for fructose versus AceK ($n = 8$ mice, two-tailed paired t -test, $P = 0.27$). Note that fructose, a caloric sugar, does not create preference, but activates osmolarity responses. **f**, Williams et al.²⁴ suggest that GPR65-expressing vagal neurons function as the nutrient sensors. We generated mice in which GCaMP6s expression was targeted to GPR65-expressing vagal neurons and examined their responses to a 10-s stimulus of 500 mM glucose or osmolarity signals (that is, 1 M each of fructose, mannose and NaCl for 60 s). Shown are normalized responses of from three different mice to the four stimuli; each trace represents a different responding neuron. Note that 500 mM glucose for 10 s does not activate GPR65 neurons. By contrast, they are activated by 60-s pulse of 1 M fructose, mannose and NaCl (see also Fig. 4). **g**, Summary histogram of GPR65 tuning profile to 10 s 500 mM glucose, and 60 s 1 M fructose, 60 s 1 M mannose and 60 s 1 M NaCl; $n = 4$ mice.



Extended Data Fig. 9 | Genetic silencing of GPR65 neurons does not affect the development of sugar preference. **a**, Global silencing of the GPR65 neurons was achieved by generating GPR65-IRES-Cre; R26-TetNT double transgenic animals expressing TetTox in GPR65 neurons. Sugar-preference graphs demonstrating the robust development of preference for sugar versus artificial sweetener for both wild-type ($n = 5$ mice, two-tailed paired t -test, $P = 0.0047$) and GPR65:TetTox mice ($n = 5$ mice, two-tailed paired t -test, $P = 0.0033$). The wild-type controls shown here are the same mice used in

Extended Data Fig. 3c, as both sets of silencing experiments were carried out as part of the same series of studies. Values are mean \pm s.e.m. **b**, Silencing of GPR65 neurons does not impair the innate attraction to sweet solutions. Shown are behavioural responses to AceK versus water and glucose versus water ($n = 5$, two-tailed paired t -test, $P = 0.0040$ for consumed volumes of AceK versus water, $P = 0.0023$ for consumed volumes of glucose versus water). Values are mean \pm s.e.m.



Extended Data Fig. 10 | Vagal neurons responding to intestinal glucose are also activated by SGLT1 agonists. **a**, Traces of vagal neurons responding to a 10-s pulse of 500 mM intestinal glucose, also challenged with a 10-s pulse of 500 mM 3-OMG. Shown are sample neurons from 2 animals. **b**, Traces of vagal neurons responding to a 10-s pulse of 500 mM intestinal glucose, also challenged with a 10-s pulse of 500 mM galactose. Shown are sample neurons from two animals for expanded time scales (from Fig. 4d). **c**, Traces of vagal neurons responding to a 10-s pulse of 500 mM intestinal glucose, also challenged with a 10-s pulse of 500 mM fructose and 500 mM mannose. Shown are sample neurons from three mice. **d**, Traces of vagal neurons responding to two consecutive 10-s pulses of 500 mM intestinal glucose, before and after

treating the intestinal segment with 8 mM phlorizin for 5 min. Note the loss of responses. **e**, Because responses, in general, show some decay during the time of the experiment (in part due to desensitizing and bleaching of the fluorescent signals), we also analysed the average decay of corresponding glucose responses in the absence of any blocker. The graphs compare the loss of responses during normal decay, and in response to the blocker. For normal decay (left), $n = 11$ neurons, Pre = 230.8 arbitrary units (a.u.), Post = 172.8 a.u.; for blocker (right), $n = 31$ neurons, Pre = 229.7 a.u., Post = 67.0 a.u. All values are mean \pm s.e.m. Scale indicates average integral of the responses to the two trials before and after inhibition.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Tucker-Davis Technologies Synapse (Version 90-39473P), MicroManager (Version 1.4), Olympus Fluoview (FV10), Arduino IDE (Version 1.8.10), MathWorks Matlab (R2019a, R2019b)

Data analysis

MathWorks Matlab (R2019b), FIJI (Version 1.52p)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data supporting the findings of this study are available upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined based similar studies in the literature and our experience. No statistical method was used to determine the sample size prior to the study.
Data exclusions	Animals in which post-hoc histological examination showed that viral targeting or the position of implanted fiber were in the incorrect location were excluded from analysis. This exclusion criteria was predetermined.
Replication	We performed multiple independent experiments as noted in the figure legends. Results were reproducible.
Randomization	Stimuli order was random, otherwise in situations as described in the manuscript where no randomization was used, the stimuli were interspersed and repeated among trials.
Blinding	Investigators were not blinded to group allocation, as data analysis was performed automatically with the same scripts executed for each experimental group.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used	anti c-Fos (Santa Cruz, SC52G, K1715, Goat, 1:500), anti c-Fos (Synaptic Systems, 226004, Guinea Pig, 1:5000)
Validation	Both antibodies has been validated extensively, e.g. by immuno-staining on mouse brain sections (Choi, et al. Cell, 146, 1004-1015, (2011), Song, et al. Science advances, 52: eaat3210, (2019)).

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Adult animals 6-24 weeks of age and from both genders were used in experiments. C57BL/6J (JAX #000664), ArcCreER (TRAP, JAX #021881), TRAP2 (JAX #030323), TRPM5 KO (JAX #013068), T1R2/T1R3 double KO (generated in house, JAX #013065/013066), Ai96 (JAX #028866), Vglut2-IRES-Cre (JAX #028863), Gpr65-IRES-Cre (JAX #029282), Penk-IRES2-Cre (JAX #025112), Ai75D (JAX #025106), R26-TenT (MGI #3839913).
Wild animals	No wild animals were used.
Field-collected samples	No field-collected samples were used.
Ethics oversight	All procedures were carried out in accordance with the US National Institutes of Health (NIH) guidelines for the care and use of laboratory animals, and were approved by the Institutional Animal Care and Use Committee at Columbia University.

Note that full information on the approval of the study protocol must also be provided in the manuscript.


Mechanisms and therapeutic implications of hypermutation in gliomas

<https://doi.org/10.1038/s41586-020-2209-9>

Received: 22 July 2019

Accepted: 4 March 2020

Published online: 15 April 2020

 Check for updates

Mehdi Touat^{1,2,3,34}✉, Yvonne Y. Li^{2,4,34}, Adam N. Boynton^{2,5}, Liam F. Spurr^{2,4}, J. Bryan Iorgulescu^{4,6}, Craig L. Bohrsen^{7,8}, Isidro Cortes-Ciriano⁹, Cristina Birzu³, Jack E. Geduldig¹, Kristine Pelton¹, Mary Jane Lim-Fat^{4,10}, Sangita Pal^{2,4}, Ruben Ferrer-Luna^{2,4,11}, Shakti H. Ramkissoon^{11,12}, Frank Dubois^{2,4}, Charlotte Bellamy¹, Naomi Currimjee⁴, Juliana Bonardi¹, Kenin Qian⁵, Patricia Ho⁵, Seth Malinowski¹, Leon Taquet¹, Robert E. Jones¹, Aniket Shetty¹³, Kin-Hoe Chow¹³, Radwa Sharaf¹¹, Dean Pavlick¹¹, Lee A. Albacker¹¹, Nadia Younan³, Capucine Baldini¹⁴, Maïté Verreault¹⁵, Marine Giry¹⁵, Errell Guillerme¹⁶, Samy Ammari^{17,18}, Frédéric Beuvon¹⁹, Karima Mokhtari²⁰, Agusti Alentorn³, Caroline Dehais³, Caroline Houillier³, Florence Laigle-Donadey³, Dimitri Psimaras³, Eudocia Q. Lee^{4,10}, Lakshmi Nayak^{4,10}, J. Ricardo McFaline-Figueroa^{4,10}, Alexandre Carpentier²¹, Philippe Cornu²¹, Laurent Capelle²¹, Bertrand Mathon²¹, Jill S. Barnholtz-Sloan²², Arnab Chakravarti²³, Wenya Linda Bi²⁴, E. Antonio Chiocca²⁴, Katie Pricola Fehnel²⁵, Sanda Alexandrescu²⁶, Susan N. Chi^{5,27}, Daphne Haas-Kogan²⁸, Tracy T. Batchelor^{4,10}, Garrett M. Frampton¹¹, Brian M. Alexander^{11,28}, Raymond Y. Huang²⁹, Azra H. Ligon⁶, Florence Coulet¹⁶, Jean-Yves Delattre^{3,30}, Khê Hoang-Xuan³, David M. Meredith^{1,6}, Sandro Santagata^{1,6,31,32}, Alex Duval³³, Marc Sanson^{3,30}, Andrew D. Cherniack^{2,4}, Patrick Y. Wen^{4,10}, David A. Reardon⁴, Aurélien Marabelle¹⁴, Peter J. Park⁷, Ahmed Idbaih³, Rameen Beroukhim^{2,4,10,35}✉, Pratiti Bandopadhyay^{2,5,27,35}✉, Franck Bielle^{20,35}✉ & Keith L. Ligon^{1,2,6,13,26,35}✉

A high tumour mutational burden (hypermutation) is observed in some gliomas^{1–5}; however, the mechanisms by which hypermutation develops and whether it predicts the response to immunotherapy are poorly understood. Here we comprehensively analyse the molecular determinants of mutational burden and signatures in 10,294 gliomas. We delineate two main pathways to hypermutation: a *de novo* pathway associated with constitutional defects in DNA polymerase and mismatch repair (MMR) genes, and a more common post-treatment pathway, associated with acquired resistance driven by MMR defects in chemotherapy-sensitive gliomas that recur after treatment with the chemotherapy drug temozolomide. Experimentally, the mutational signature of post-treatment hypermutated gliomas was recapitulated by temozolomide-induced damage in cells with MMR deficiency. MMR-deficient gliomas were characterized by a lack of prominent T cell infiltrates, extensive intratumoral heterogeneity, poor patient survival and a low rate of response to PD-1 blockade. Moreover, although bulk analyses did not detect microsatellite instability in MMR-deficient gliomas, single-cell whole-genome sequencing analysis of post-treatment hypermutated glioma cells identified microsatellite mutations. These results show that chemotherapy can drive the acquisition of hypermutated populations without promoting a response to PD-1 blockade and supports the diagnostic use of mutational burden and signatures in cancer.

Identifying genomic markers of response to immune checkpoint blockade (for example, PD-1 blockade) may benefit cancer patients by providing predictive biomarkers for patient stratification and identifying resistance mechanisms for therapeutic targeting. Gliomas typically have a low tumour mutational burden (TMB) and a highly immunosuppressive microenvironment—two features associated with immunotherapy resistance. Nevertheless, recent work has suggested that a subset of patients with high-TMB (hypermutated) gliomas might benefit from PD-1 blockade⁶. Although consistent with

data from other cancers^{7–9}, these initial observations were derived from unique disease contexts such as constitutional DNA mismatch-repair (MMR) deficiency syndrome⁶. Therefore, the extent to which glioma patients at large will benefit from this approach is unknown. While large amounts of genomic data on gliomas exist^{2,4,5,10,11,12}, our understanding of the clinical landscape of hypermutation and the mechanisms that underlie its development remain unclear. Hypermutation is rare in newly-diagnosed gliomas (*de novo* hypermutation), but common in tumours that have recurred after the use of alkylating

A list of affiliations appears at the end of the paper.

agents (post-treatment hypermutation)^{4,5,10,11}. Given that gliomas exhibit substantial inter-patient and intra-tumoral genomic variation^{10,11,12}, it remains to be determined whether molecular biomarkers (for example, *IDH1* or *IDH2* (hereafter *IDH1/2*) mutations) reliably predict the development of hypermutation or response to immunotherapy.

An association between hypermutation and MMR mutations has been observed in gliomas^{1–4,13}, but most of the reported MMR mutations were not functionally characterized, and their role in causing hypermutation is unclear. Other studies have suggested that alkylating agents such as temozolomide are the direct cause of hypermutation³. This was supported by the discovery of a mutational signature (single base substitution (SBS) signature 11) characterized by the accumulation of G:C>A:T transitions at non-CpG sites in hypermutated gliomas after exposure to alkylating agents¹⁴. However, the fact that hypermutation is undetectable in most gliomas that recur after temozolomide treatment challenges this notion^{4,5}. Furthermore, it remains unclear whether this mutational pattern enhances tumour immunogenicity and renders gliomas responsive to PD-1 blockade. Not all hypermutated cancers respond to such treatments^{7–9}; a more accurate characterization of the phenotypic and molecular features of hypermutated gliomas therefore would help clinicians to manage such patients more effectively.

Mutational burden and signatures in gliomas

Previous studies included too few hypermutated gliomas to characterize the landscape of hypermutation in gliomas^{1–5}. We therefore created a cohort of sufficient scale ($n = 10,294$) and subtype diversity by leveraging large datasets generated from clinical sequencing panels (DFCI-Profile, MSKCC-IMPACT and FMI)^{15–17}. All samples from patients with a histopathological diagnosis of glioma were included and classified into molecular subgroups according to histopathology, mutational status of *IDH1/2*, and whole-arm co-deletion of chromosomes 1p and 19q (1p/19q co-deletion) (Extended Data Fig. 1, Supplementary Tables 1, 2). We quantified the TMB of all samples (median 2.6 mutations (mut.) per Mb (range 0.0–781.3)), established thresholds for hypermutation by examining the distribution of TMB (Extended Data Fig. 2)^{17,18}, and identified 558 (5.4%) hypermutated gliomas (median TMB 50.8 mut. per Mb (8.8–781.3)) for further analysis.

Using samples with detailed clinical annotation (DFCI-Profile), we found that the prevalence of hypermutation varied between and within subgroups (Fig. 1a, b, Extended Data Fig. 3a, b, Supplementary Table 3). Hypermutation was detected almost exclusively in diffuse gliomas (99.1% of hypermutated samples) with high-grade histology (95.6%) and was more prevalent in recurrent tumours (16.6% versus 2.0% in newly diagnosed tumours; Fisher's exact test, $P < 10^{-15}$) (Fig. 1b). In samples of recurrent tumours, hypermutation was associated with markers of response to alkylating agents, including *IDH1/2* mutation (hypermutation in 1.4% of newly diagnosed versus 25.4% of post-treatment *IDH1/2*-mutant tumours, Fisher's exact test, $P = 2.0 \times 10^{-13}$), 1p/19q co-deletion (0.0% versus 33.8%, $P = 7.3 \times 10^{-11}$), and *MGMT* promoter methylation (2.4% versus 24.2%, $P = 9.0 \times 10^{-12}$). The effect of *IDH1/2* mutation was confirmed only in *MGMT*-methylated tumours (Extended Data Fig. 3c). These findings suggest that selective pressure from therapy may elicit progression towards hypermutation.

The standard treatment for gliomas includes surgery, radiation and chemotherapy with alkylating agents^{19,20}. To assess the role of each of these in the development of hypermutation, we analysed associations between TMB and detailed patterns of treatment in 356 recurrent gliomas. Hypermutation was associated with prior treatment with temozolomide (Fisher's exact test, $P < 10^{-15}$) in a dose-dependent manner (Fig. 1b, Extended Data Fig. 3d, e), but not with radiation ($P = 0.88$) or nitrosoureas ($P = 0.78$). Among recurrent tumours from patients who had received only one adjuvant treatment modality, TMB was increased only in temozolomide-treated samples (median 16.32 (interquartile range (IQR) 6.95–70.32) versus 6.08 (3.80–7.97) with surgery

only, $P = 4.0 \times 10^{-7}$; Extended Data Fig. 3f). Of note, the prevalence of hypermutation in post-temozolomide samples correlated with the chemosensitivity of the primary, molecularly defined tumour type (1p/19q co-deleted oligodendrogliomas (59.5%) > *IDH1/2*-mutant astrocytomas (30.2%) > *MGMT*-methylated *IDH1/2* wild-type glioblastomas (23.1%) > *MGMT*-unmethylated *IDH1/2* wild-type glioblastomas (5.6%); $P = 3.8 \times 10^{-7}$; Fig. 1b). We observed a similar pattern in the FMI validation dataset (Extended Data Fig. 3g–i).

The systematic analysis of somatic mutation patterns by genome sequencing has identified a variety of mutation signatures in human cancer which are driven by known and unknown DNA damage and repair processes¹⁴. We examined the contributions of 30 previously reported signatures (COSMIC signatures v2) within our cohort to investigate the biological processes that cause hypermutation in gliomas. We first validated that mutational signatures can be predicted using large targeted panel sequencing in hypermutated samples (Extended Data Figs. 4, 5a–c). The majority of de novo hypermutated gliomas harboured mutational signatures associated with defects in the MMR pathway (COSMIC signatures 6, 15, 26 and 14) or the DNA polymerase POLE (10 and 14)¹⁴ (69% and 35% of samples, respectively; Extended Data Fig. 5d, e), implying that constitutional deficiency in MMR or POLE was likely to be the underlying genetic cause of hypermutation. By contrast, 98% of post-treatment hypermutated gliomas showed a mutational signature that has been previously associated with temozolomide exposure (signature 11). We also identified two distinct mutational signatures that were highly correlated with mutational signature 11 (Extended Data Fig. 5b, c) including a previously undescribed signature (S2) associated with 1p/19q co-deletion and lack of prior radiation therapy. Finally, half of the samples with a dominant signature 11 showed a co-existing minor MMR- or POLE-deficiency signature component (Extended Data Fig. 5e), suggesting that defective DNA repair and mutagen exposure cooperate to drive hypermutation in recurrent gliomas.

Molecular drivers of hypermutation

Only a subset of temozolomide-treated samples (58 of 225, 25.8%) showed evidence of hypermutation, suggesting that additional factors are required for its development. Although MMR defects have been consistently observed in hypermutated gliomas^{1–4,13}, their co-occurrence with high TMB did not enable prior studies to determine the degree to which MMR mutations represent passenger versus hypermutation-causing driver events. We systematically characterized mutations and copy number variants (CNVs; Supplementary Figs. 1, 2) to identify hypermutation drivers using an unbiased approach that controlled for the increased incidence of passenger mutations associated with hypermutation²¹. In the merged DFCI-Profile/MSKCC-IMPACT dataset, 36 genes were significantly enriched (q value < 0.01) in hypermutated tumours (Fig. 2a). Collectively, MMR mutations stood out among the most enriched (91.2% versus 4.9% in non-hypermutated samples, $q < 1.6 \times 10^{-15}$), and mutations in *MSH6* showed the highest enrichment (43.0% versus 1.2%, $q = 3.3 \times 10^{-7}$) (Extended Data Figs. 3j–l, 6a, b). MMR-variant allele frequencies (VAFs) and cancer cell fractions (CCFs) in gliomas were most similar to those in MMR-deficient colorectal (CRC) or endometrial cancers and were higher than in MMR-proficient hypermutated cancers (Extended Data Fig. 6c, d). Some MMR variants in post-treatment hypermutated samples matched the canonical signature 11 sequence context (Extended Data Fig. 5f), suggesting that a subset of these variants are likely to have been caused by temozolomide treatment.

As most MMR variants lacked functional annotation, we next integrated sequencing data with immunohistochemistry for protein loss (Extended Data Fig. 6e). Overall, results from both assays were concordant, consistent with MMR mutations leading to loss of function. In rare samples that lacked MMR variants, signature analysis and MMR immunohistochemistry revealed evidence for MMR deficiency,

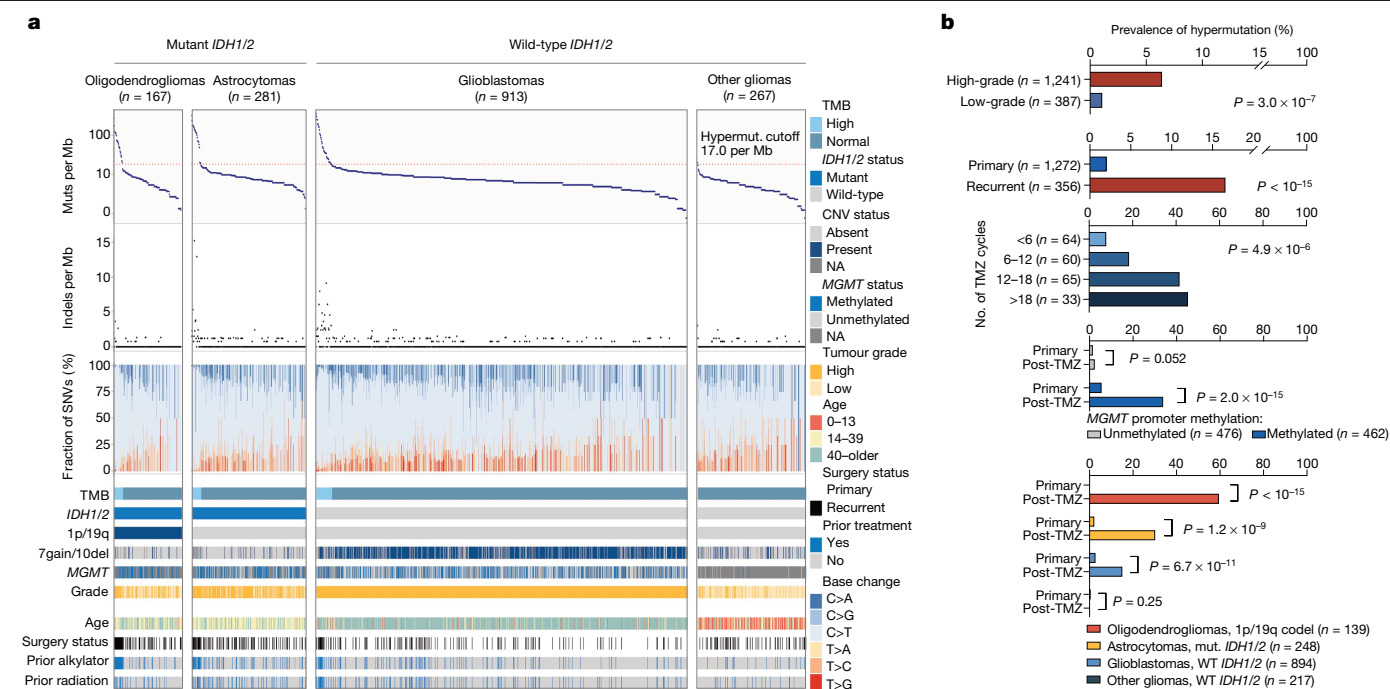


Fig. 1 | TMB and mutational signature analysis reveals clinically distinct subgroups of hypermutated gliomas. a, Integrated analysis of the DFCI-Profile dataset ($n = 1,628$ gliomas) depicting TMB, indels at homopolymer regions, and the single nucleotide variant (SNV) mutation spectrum in each tumour according to molecular status of *IDH1/2*, 1p/19q co-deletion, chromosome 7 gain and/or chromosome 10 deletion (7gain/10del), *MGMT*

promoter methylation, histological grade, age at initial diagnosis, and prior treatment. Red line denotes high TMB (≥ 17.0 mut. per Mb). **b**, Prevalence of hypermutation in the DFCI-Profile dataset. Chi-squared test and two-sided Fisher's exact test. NA, not available; TMZ, temozolomide; WT, wild-type; mut, mutant; code, co-deleted.

suggesting that these samples harboured underlying MMR defects that could not be identified by sequencing (for example, promoter methylation). We identified several MMR mutational hotspots (Extended Data Fig. 6f, Supplementary Table 4), including a recurrent *MSH6* mutation (p.T1219I, in 7.4% of hypermutated tumours) that has been previously identified in Lynch syndrome and shown to exert a dominant-negative effect without affecting protein expression^{22,23} (Extended Data Fig. 6g, h).

Immunohistochemistry on an independent cohort of 213 recurrent post-alkylator gliomas further validated these findings (Supplementary Table 2). MMR protein expression was lost in 22 post-treatment samples, and this loss was associated with *IDH1/2* mutations (20% mutant versus 2% wild-type; Fisher's exact test, $P = 8.0 \times 10^{-6}$) (Extended Data Fig. 7a, b). Sequencing of samples with MMR protein loss confirmed hypermutation, with MMR mutations in 18 of 19 (94.7%) of these samples. Subclonal loss of MMR proteins (that is, protein retained in more than 20% of tumour cells) was more common in post-treatment than de novo hypermutated gliomas (12 of 46 (26.1%) versus 0 of 16 (0.0%), $P = 0.03$) (Extended Data Fig. 7c–f).

We next assessed the relationship between MMR deficiency and acquired chemotherapy resistance. Because hypermutation and MMR defects were almost exclusively seen after temozolomide treatment, we hypothesized that nitrosoureas and temozolomide might not show complete cross-resistance. Analysis of temozolomide sensitivity in 30 cell lines derived from patients with glioma (patient-derived cell lines, PDCLs), including four derived from MMR-deficient gliomas (Extended Data Fig. 8a–c), showed that all native MMR-deficient PDCLs had striking temozolomide resistance compared to MMR-proficient PDCLs (6.46- and 1.35-fold increase in median area under the curve (AUC) versus MMR-proficient–MGMT-deficient and MMR-proficient–MGMT-proficient PDCLs, respectively) (Fig. 2b, Extended Data Fig. 8d–f). We next treated native and engineered isogenic MMR-knockout

glioma models with temozolomide or the nitrosourea lomustine (CCNU), a chloroethylating alkylating agent that generates DNA interstrand crosslinks and double-strand breaks (Fig. 2c, Extended Data Fig. 8g–i). All MMR-deficient models were resistant to temozolomide and sensitive to CCNU, consistent with the lack of hypermutation in samples from nitrosourea-treated patients²⁴ (Extended Data Fig. 3f).

Mismatch repair deficiency and signature 11

Our analyses indicated that MMR deficiency together with temozolomide exposure might cause signature 11, as opposed to it being a 'pure' temozolomide signature. To test this idea, we exposed isogenic models of MMR deficiency to temozolomide (Extended Data Fig. 9a, b). After treatment with temozolomide, MMR-deficient PDCLs developed hypermutation with signature 11, whereas MMR-proficient controls (expressing sgGFP) did not (Fig. 2d). We then chronically treated temozolomide-sensitive glioblastoma xenografts (PDXs) with temozolomide until resistance was acquired (Fig. 2e, Extended Data Fig. 9c, d). These tumours developed hypermutation with signature 11 (Fig. 2f, Extended Data Fig. 9e) and shared four unique variants; the dominant-negative *MSH6* hotspot mutation (p.T1219I) and three non-coding variants (Fig. 2g), consistent with the theory that the *MSH6* mutation drives both hypermutation and acquired temozolomide resistance (Extended Data Fig. 9f).

Collectively, these findings show that temozolomide exerts a previously underappreciated selective pressure in favour of MMR-deficient cells, which are resistant to temozolomide. Exposing MMR-deficient cells to temozolomide induces hypermutation with signature 11 by causing DNA damage in the absence of functional MMR. Therefore, hypermutation with signature 11 represents a potential biomarker for MMR deficiency and temozolomide resistance in gliomas (Extended Data Fig. 9g).

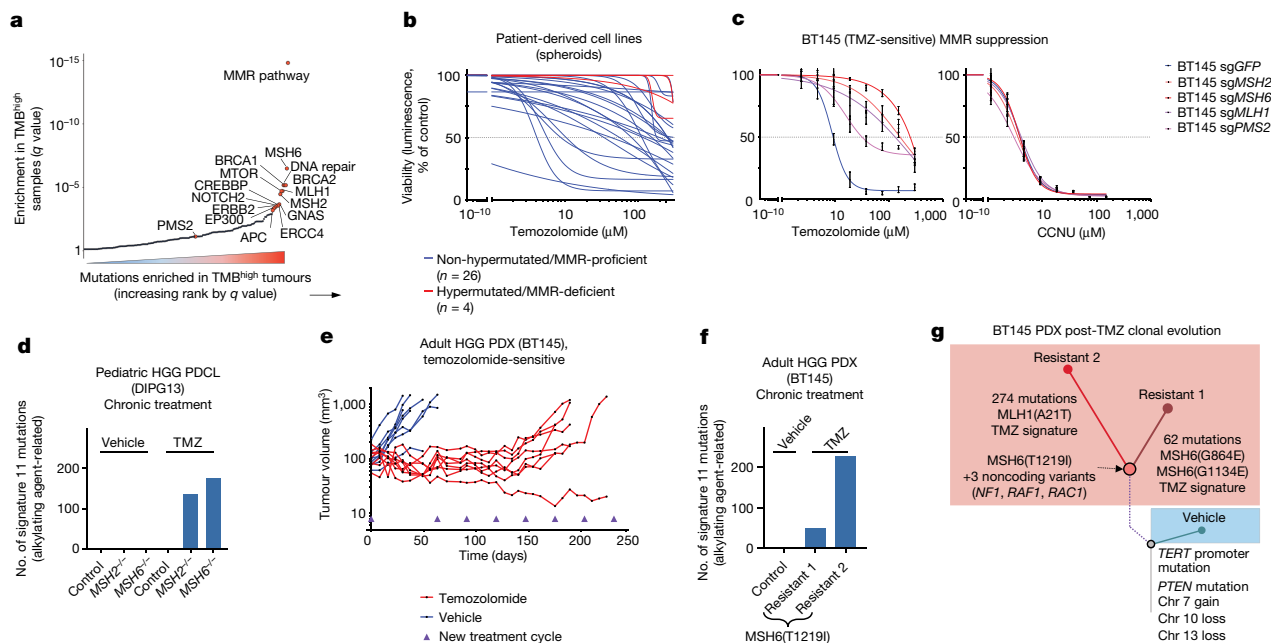


Fig. 2 | MMR deficiency drives hypermutation and chemotherapy resistance in gliomas. a, Mutated genes and pathways enriched in hypermutated gliomas in the merged DFCI-Profile/MSKCC-IMPACT dataset ($n = 2,173$) using a permutation test to control for random mutation rate in the setting of hypermutability. **b**, Response to temozolomide across a panel ($n = 30$) of native spheroid glioma PDCLs (blue, MMR-proficient; red, MMR-deficient). Dose-response curves were calculated using mean surviving fractions from three independent assays. **c**, Response to temozolomide and CCNU in the glioblastoma PDCL BT145 following knockout of *MSH2*, *MSH6*, *MLH1* or *PMS2* by CRISPR-Cas9. Dose-response curves were calculated using mean surviving fractions from three independent assays (mean \pm s.e.m.). **d**, Number of signature 11 variants after chronic temozolomide treatment of the PDCL DIPG13 with *MSH2* or *MSH6* knockout by CRISPR-Cas9. Mutational signatures could not be called in

the vehicle-treated samples (too few variants). **e**, Tumour volume ($n = 8$ mice per group) during treatment with vehicle (blue) or temozolomide (red) in BT145 patient-derived xenografts (PDXs). **f**, Number of signature 11 variants found after chronic temozolomide exposure in BT145 PDXs. Mutational signatures could not be called in the vehicle-treated tumours (too few variants). **g**, Schematic representation of BT145 PDXs clonal evolution under temozolomide exposure. Two independent secondary resistant tumours (Resistant 1 and 2) and one vehicle-treated tumour are represented. Resistant tumours have four private variants that were not detected in the vehicle-treated tumour: an *MSH6*(T1219I) mutation (VAF 0.27 and 0.37 for resistant 1 and 2, respectively), and three non-coding variants of *NF1* (VAF 1.0 and 0.99), *RAC1* (VAF 0.86 and 0.86) and *RAF1* (0.44 and 0.56). HGG, high-grade glioma; Chr, chromosome.

Characteristics of MMR-deficient gliomas

MMR deficiency recently emerged as an indicator of response to PD-1 blockade in patients with cancer^{8,25}, leading to the first tissue-agnostic cancer-drug approval by the US Food and Drug Administration for use of the PD-1 blocker pembrolizumab in patients with MMR-deficient cancers. However, in CRCs and some other cancers, MMR inactivation occurs early in tumour progression, whereas in post-treatment gliomas it arises late. Gliomas might therefore differ from other cancers on which the approval was based and these differences might influence immune recognition of tumours and the response to immunotherapy.

To test this hypothesis, we first assessed the outcome of hypermutated gliomas. In CRC, MMR deficiency is associated with improved outcomes. By contrast, among patients with recurrent glioma, we observed worse survival in both hypermutated high-grade 1p/19q co-deleted oligodendrogliomas (median overall survival (OS) 96.5 months (95% confidence interval (CI) 20.8–NA (not applicable)) versus 137.2 months (95% CI 41.8–NA) in non-hypermutated tumours, $P = 0.0009$, two-sided log-rank test) and *IDH1/2*-mutant astrocytomas (median OS 15.7 months (95% CI 12.9–18.3) versus 21.5 months (95% CI 19.2–29.8), $P = 0.0015$) (Fig. 3a, Extended Data Fig. 10a–c). We observed a similar trend in *IDH1/2* wild-type glioblastomas ($P = 0.0809$). The finding of poor survival in recurrent hypermutated gliomas remained significant in multivariable analysis (hazard ratio 2.16 (95% CI 1.38–3.38), $P = 0.0008$; Supplementary Table 5).

The current hypothesis behind the response of MMR-deficient CRCs to PD-1 blockade is based on their increased neoantigen burden and immune infiltration. We therefore assessed the association between

MMR deficiency and T-cell infiltration in gliomas ($n = 43$) and CRCs ($n = 19$). As expected, MMR-deficient CRCs exhibited significantly more infiltrating T-cells than their MMR-proficient counterparts (Fig. 3b). By contrast, both MMR-deficient and MMR-proficient glioma samples lacked significant T-cell infiltrates (Fig. 3c).

We next assessed whether the neoantigen burden was lower in MMR-deficient gliomas than in other hypermutated cancers using samples from the GENIE and TCGA datasets ($n = 1,748$ and 699 hypermutated cancers, respectively). As neoantigen prediction was not feasible using panel sequencing data, we used the nonsynonymous mutational burden as a surrogate measure. This showed that both de novo and post-treatment MMR-deficient gliomas had an increase in their nonsynonymous mutational burden, when compared to non-hypermutated gliomas, and the glioma nonsynonymous mutational burden was similar to other hypermutated cancers (Fig. 3d, Extended Data Fig. 11a, b, Supplementary Table 6). This finding suggested that the total number of neoantigens is unlikely to explain the differences in immune response between gliomas and other hypermutated cancers.

Recent data suggest that, among mutations associated with MMR deficiency, small insertions and deletions (indels) at homopolymers (microsatellites)—which accumulate in MMR-deficient cells and can cause frameshift mutations—are crucial for producing ‘high-quality’ neoantigens that are recognized by immune cells^{26,28}. Unexpectedly, although the high TMB was associated with an increased homopolymer indel burden in MMR-deficient CRCs, this association was not found in MMR-deficient gliomas (de novo hypermutated gliomas showed a modest increase; Fig. 3d, Extended Data Fig. 11c). This was validated using testing for microsatellite instability (MSI), a clinical biomarker for MMR

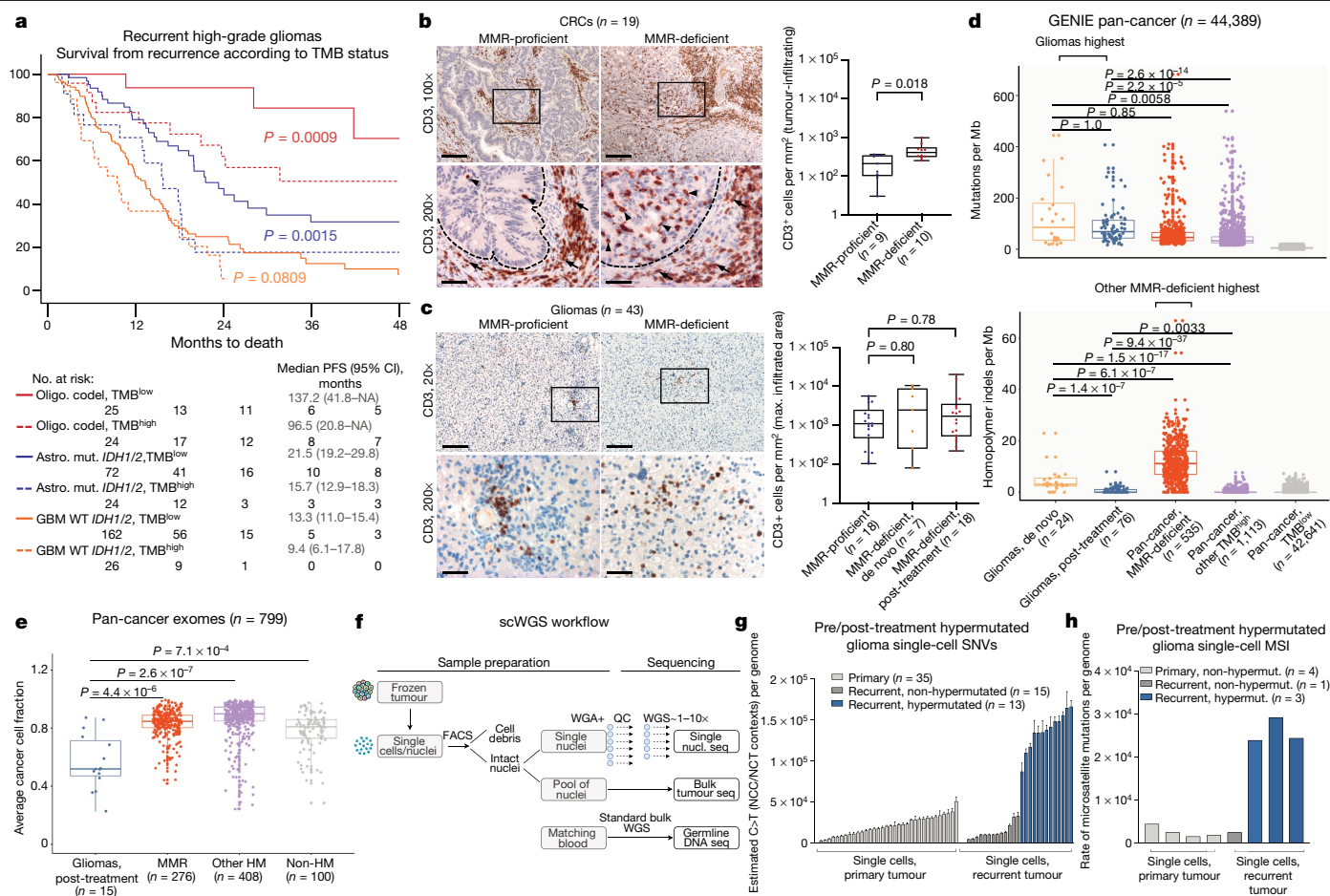


Fig. 3 | Hypermutated and MMR-deficient gliomas harbour unique phenotypic and molecular characteristics including poor outcome and lack of MSI in bulk sequencing.

a, Survival of patients with recurrent high-grade glioma from the time of sample collection according to histomolecular group and TMB status ($n = 333$ recurrent samples; 238 from DFCI-Profile, 95 from MSKCC-IMPACT). Two-sided log-rank test. **b**, Quantification of tumour-infiltrating CD3-positive T-cells in CRC samples ($n = 19$). Left, representative low- and high-magnification images of CD3 immunolabelling (brown; intraepithelial lymphocytes, black arrowheads; stromal lymphocytes, black arrows) and nuclear counterstaining (blue). Dashed lines, border between tumour and stroma. Only intraepithelial lymphocytes were quantified. Scale bars; 100 μ m (100 \times), 50 μ m (200 \times). Right: boxes, quartiles; centre lines, median ratio for each group; whiskers, absolute range. Two-sided Wilcoxon rank-sum test. **c**, Quantification of tumour-infiltrating CD3-positive T-cells in gliomas according to their MMR status ($n = 43$). For each group, three areas with the maximal CD3 infiltration were selected for quantification (representative images, left). Scale bars: 500 μ m (20 \times), 50 μ m (200 \times). Right: boxes, quartiles; centre lines, median ratio for each group; whiskers, absolute range. Kruskal–Wallis test and Dunn’s multiple comparison test. **d**, TMB (top) and homopolymer indel burden

(bottom) in hypermutated gliomas compared with other hypermutated cancers from the GENIE dataset. Tukey’s boxplots are shown. Two-sided Wilcoxon rank-sum test with Bonferroni correction. **e**, Pan-cancer analysis of cancer cell fractions in hypermutated gliomas (post-treatment) compared with other hypermutated cancers from the TCGA and ref. ⁴ exome datasets ($n = 798$). One hundred non-hypermutated samples from the TCGA were randomly selected as controls. Boxes, quartiles; centre lines, median ratio for each group; whiskers, absolute range excluding outliers. Two-sided Wilcoxon rank-sum test with Bonferroni correction. **f**, Workflow for scWGS and bulk tumour DNA sequencing. **g**, Single-cell sequencing estimate of the number of G:C>A:T transitions at NCC and NCT trinucleotide contexts in 63 cells from a glioblastoma patient with post-temozolomide hypermutation using 1 \times scWGS sequencing. Error bars show 95% CI. The absolute computed purity was 0.66 for the primary tumour sample and 0.47 for the recurrent tumour sample in the bulk sequencing. **h**, Single-cell sequencing estimate of microsatellite mutation rate in eight cells from a patient with glioblastoma with post-temozolomide hypermutation. Eight cells were analysed for the presence of MSI using 10 \times scWGS sequencing. WGA, whole genome amplification; QC, quality control; nucl, nuclei; seq, sequencing.

deficiency. Whereas MSI was identified in all MMR-deficient CRCs, all tested gliomas with MMR protein loss ($n = 15$) were microsatellite-stable (MSS) (Extended Data Figs. 7d–f, 11d).

We hypothesized that, in hypermutated gliomas, more of the homopolymer indels are subclonal and below the detection limits of bulk sequencing, relative to other MMR-deficient cancers. Indeed, analysis of CCFs indicated that hypermutated gliomas contained a greater burden of subclonal variants than did other hypermutated cancers (Fig. 3e, Extended Data Fig. 11e–h). We therefore performed single-cell whole-genome DNA sequencing (scWGS) of 28 cells from a hypermutated, post-temozolomide glioblastoma with an MSH6(T1219I)

mutation, and compared these to 35 non-hypermutated cells from the matched pre-treatment sample (Fig. 3f, Extended Data Fig. 11i–k). In the post-temozolomide sample, 13 of 28 cells (46.4%) were hypermutated with signature 11 (Fig. 3g, Extended Data Fig. 11l). Strikingly, whereas this tumour harboured only a minor increase in its homopolymer indel burden at the bulk level (0.49 versus 0.0 per Mb), the scWGS analysis showed a ninefold increase in microsatellite mutations in all hypermutated cells (Fig. 3h). This suggested that glioma cells with an MSH6(T1219I) variant harbour a subtle MSI phenotype that is not revealed by standard bulk sequencing or clinical MSI assays (Extended Data Fig. 11m).

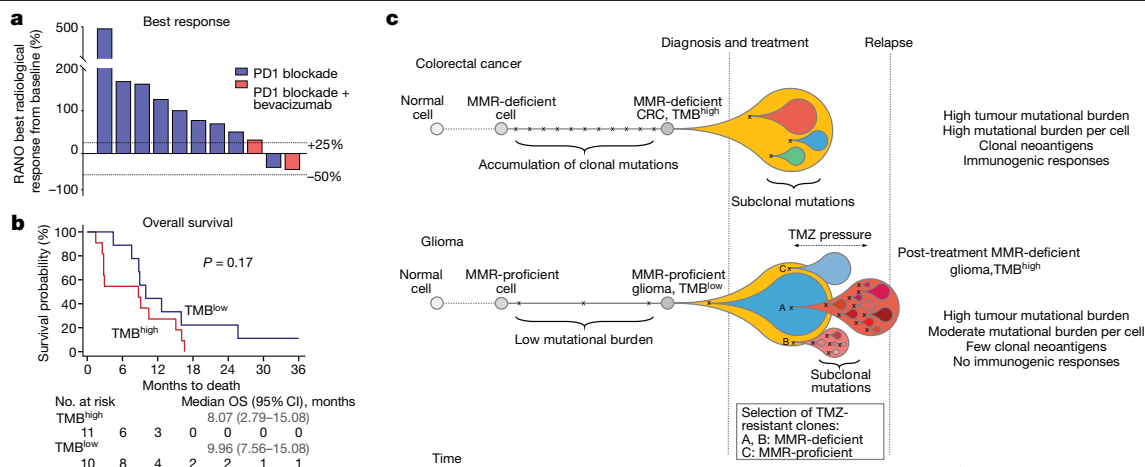


Fig. 4 | Treatment of hypermutated gliomas with PD-1 blockade. **a**, **b** Best radiological response (**a**, measured as the best change in the sum of the products of perpendicular diameters of target lesions), and overall survival (**b**) of 11 patients with hypermutated and MMR-deficient gliomas who were treated with PD-1 blockade. A cohort of patients with non-hypermutated gliomas who were treated with PD-1 blockade is depicted as control ($n = 10$, best matches according to diagnosis, primary versus recurrent status, and prior treatments). Two-sided log-rank test. **c**, Proposed model explaining differential response to PD-1 blockade in MMR-deficient CRCs and gliomas. In CRCs (top), MMR deficiency is acquired early in pre-cancerous cells, creating mutations and indels at homopolymer regions. Over time, clonal neoantigens of both types emerge and

strong immune infiltrates are seen at diagnosis. Treatment with anti-PD-1 results in expansion of T cells that recognize these clonal neoantigens and substantial antitumour responses. In gliomas (bottom), few mutations are acquired early during tumorigenesis in the majority of tumours. Temozolomide drives the expansion of cells with MMR deficiency and late accumulation of random temozolomide-induced mutations. Ineffective antitumour responses may result from poor neoantigen quality (high burden of missense mutations versus frameshift-producing indels) and high subclonality associated with an immunosuppressive microenvironment. In some tumours, MMR-proficient subclones that have acquired therapy resistance through other pathways can co-exist with MMR-deficient subclones, giving rise to a mixed phenotype.

PD-1 blockade in MMR-deficient gliomas

As hypermutation in gliomas that acquire MMR deficiency tends to be subclonal and does not generate optimal antitumour T-cell responses, we hypothesized that these tumours might not have high response rates to PD-1 blockade. We performed a retrospective institutional review of patients treated with PD-1 pathway blockade for which the TMB at treatment initiation was available ($n = 210$). This identified 11 patients with MMR-deficient glioma (5 de novo, 6 post-treatment) who were treated with PD-1 blockade for a median of 42 days (range 13–145; Supplementary Table 7). Nine (81.8%) had disease progression as their best response (Fig. 4a), and the median progression-free survival (PFS) and OS were 1.38 months (95% CI 0.95–2.69) and 8.7 months (95% CI 2.79–15.08), which were not significantly different from the data for matched patients with non-hypermutated glioma (PFS 1.87 months (95% CI 1.28–2.92), OS 9.96 months (95% CI 7.56–15.08); Fig. 4b, Extended Data Fig. 10d).

Because our prior analyses indicated that patients with hypermutated gliomas might have reduced survival, we used a second set of historical controls to compare the outcome of hypermutated gliomas treated with PD-1 blockade versus other systemic agents (Supplementary Table 7). Unexpectedly, we observed a longer median OS for patients treated with other systemic agents when compared to those treated with PD-1 blockade (16.10 months (95% CI 3.98–22.21) versus 8.07 (95% CI 2.79–15.08.21); $P = 0.02$, two-sided log-rank test; Extended Data Fig. 10e, f, Supplementary Table 8). In one patient with hypermutated glioma that showed rapid imaging changes, histopathologic analysis of samples taken before and after treatment with PD-1 blockade showed highly proliferative tumour in both samples, with no significant evidence of pathologic response or increase in immune infiltrates after PD-1 blockade (Extended Data Fig. 10g).

DISCUSSION

Collectively, these results support a model in which differences in the mutation landscape and antigen clonality of hypermutated gliomas relative to other hypermutated cancers markedly affect the

response to immunotherapy (Fig. 4c) and may explain the lack of both recognition of MMR-deficient glioma cells by the host immune system and response to PD-1 blockade, compared to other MMR-deficient cancers^{8,25}. A key difference is that MMR-deficient gliomas lack detectable MSI by standard assays, similar to data from patients with constitutional MMR deficiency syndromes³⁰. Our scWGS analyses suggest that this discordance might be due to intratumour heterogeneity and a lack of sufficient evolutionary time to select clonal MSI populations. Mechanistically, selective pressure exerted by temozolomide drives the late evolution of MMR-deficient subclones, which further accumulate temozolomide-induced mutations in individual cells. In line with previous data, therapy-induced single nucleotide variant mutations might not elicit effective antitumour responses, possibly because of the quality (missense mutations versus frameshift-producing indels) or subclonal nature of their associated neoantigens^{8,27–29}. However, future evaluation of longer treatment exposure or combinatorial strategies is warranted to determine whether checkpoint blockade can be effective in this or other selected populations (for example, individuals with newly diagnosed MMR- or POLE-deficient gliomas)⁶.

We have presented evidence that recurrent defects in the MMR pathway drive hypermutation and acquired temozolomide resistance in chemotherapy-sensitive gliomas. Although it is difficult to determine the origin of MMR deficiency by sequence context alone in individual post-treatment samples, our data suggests that some MMR variants are likely to be caused by temozolomide. However, as acquired MMR deficiency occurs in the most temozolomide-sensitive tumours, it is not clear whether the acquired MMR deficiency outweighs the positive effects of temozolomide in gliomas. Our finding that MMR-deficient cells retain sensitivity to CCNU supports the hypothesis that hypermutation reduces cellular fitness and tolerance to DNA-damaging agents other than temozolomide. These alternatives are of interest in light of recent evidence showing that the addition of CCNU to chemoradiation improves the outcome of patients with *MGMT*-methylated glioblastomas³¹. Future studies are warranted to address the possibility that upfront temozolomide with CCNU may attenuate the process of post-treatment hypermutation. Furthermore, mechanisms of

resistance to temozolomide that are not associated with hypermutation will need to be addressed.

Finally, our data indicate that the absence of an immune response in gliomas is likely to result from several aspects of immunosuppression in the brain that require further characterization. Approaches that increase infiltration by cytotoxic lymphocytes into the glioma micro-environment will probably be required to improve immunotherapy response. Our data also suggest a change in practice whereby repeated biopsies and sequencing to identify progression and hypermutation could inform prognosis and guide therapeutic management.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2209-9>.

- Hunter, C. et al. A hypermutation phenotype and somatic MSH6 mutations in recurrent human malignant gliomas after alkylator chemotherapy. *Cancer Res.* **66**, 3987–3991 (2006).
- Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
- Johnson, B. E. et al. Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189–193 (2014).
- Wang, J. et al. Clonal evolution of glioblastoma under therapy. *Nat. Genet.* **48**, 768–776 (2016).
- Barthel, F. P. et al. Longitudinal molecular trajectories of diffuse glioma in adults. *Nature* **576**, 112–120 (2019).
- Bouffet, E. et al. Immune checkpoint inhibition for hypermutant glioblastoma multiforme resulting from germline biallelic mismatch repair deficiency. *J. Clin. Oncol.* **34**, 2206–2211 (2016).
- Rizvi, N. A. et al. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science* **348**, 124–128 (2015).
- Le, D. T. et al. PD-1 blockade in tumors with mismatch-repair deficiency. *N. Engl. J. Med.* **372**, 2509–2520 (2015).
- McGranahan, N. et al. Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science* **351**, 1463–1469 (2016).
- Brennan, C. W. et al. The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477 (2013).
- Brat, D. J. et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *N. Engl. J. Med.* **372**, 2481–2498 (2015).
- Louis, D. N. et al. World Health Organization Histological Classification of Tumours of the Central Nervous System (ed. 2) (International Agency for Research on Cancer, 2016).
- Cahill, D. P. et al. Loss of the mismatch repair protein MSH6 in human glioblastomas is associated with tumor progression during temozolomide treatment. *Clin. Cancer Res.* **13**, 2038–2045 (2007).
- Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
- Frampton, G. M. et al. Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. *Nat. Biotechnol.* **31**, 1023–1031 (2013).
- Sholl, L. M. et al. Institutional implementation of clinical tumor profiling on an unselected cancer population. *JCI Insight* **1**, e87062 (2016).
- Zehir, A. et al. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nat. Med.* **23**, 703–713 (2017).
- Campbell, B. B. et al. Comprehensive analysis of hypermutation in human cancer. *Cell* **171**, 1042–1056.e10 (2017).
- Stupp, R. et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N. Engl. J. Med.* **352**, 987–996 (2005).
- van den Bent, M. J. et al. Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: long-term follow-up of EORTC brain tumor group study 26951. *J. Clin. Oncol.* **31**, 344–350 (2013).
- Kim, J. et al. Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. *Nat. Genet.* **48**, 600–606 (2016).
- Berends, M. J. et al. Molecular and clinical characteristics of MSH6 variants: an analysis of 25 index carriers of a germline variant. *Am. J. Hum. Genet.* **70**, 26–37 (2002).
- Yang, G. et al. Dominant effects of an Msh6 missense mutation on DNA repair and cancer susceptibility. *Cancer Cell* **6**, 139–150 (2004).
- Ollier, E. et al. Analysis of temozolomide resistance in low-grade gliomas using a mechanistic mathematical model. *Fundam. Clin. Pharmacol.* **31**, 347–358 (2017).
- Marabelle, A. et al. Efficacy of pembrolizumab in patients with noncolorectal high microsatellite instability/mismatch repair-deficient cancer: results from the phase II KEYNOTE-158 study. *J. Clin. Oncol.* **38**, 1–10 (2020).
- Germano, G. et al. Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. *Nature* **552**, 116–120 (2017).
- Mandal, R. et al. Genetic diversity of tumors with mismatch repair deficiency influences anti-PD-1 immunotherapy response. *Science* **364**, 485–491 (2019).
- Rosenthal, R. et al. Neoantigen-directed immune escape in lung cancer evolution. *Nature* **567**, 479–485 (2019).
- Gejman, R. S. et al. Rejection of immunogenic tumor clones is limited by clonal fraction. *eLife* **7**, e41090 (2018).
- Gylling, A. H. et al. Differential cancer predisposition in Lynch syndrome: insights from molecular analysis of brain and urinary tract tumors. *Carcinogenesis* **29**, 1351–1359 (2008).
- Herrlinger, U. et al. Lomustine-temozolomide combination therapy versus standard temozolomide therapy in patients with newly diagnosed glioblastoma with methylated MGMT promoter (CeTeG/NOA-09): a randomised, open-label, phase 3 trial. *Lancet* **393**, 678–688 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

¹Department of Oncologic Pathology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. ²Broad Institute of Harvard and MIT, Cambridge, MA, USA. ³Sorbonne Université, Inserm, CNRS, UMR S 1127, Institut du Cerveau et de la Moelle épinière, ICM, AP-HP, Hôpitaux Universitaires La Pitié Salpêtrière - Charles Foix, Service de Neurologie 2-Mazarin, Paris, France. ⁴Department of Medical Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. ⁵Dana-Farber/Boston Children's Cancer and Blood Disorders Center, Boston, MA, USA. ⁶Department of Pathology, Brigham & Women's Hospital, Boston, Harvard Medical School, MA, USA. ⁷Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA. ⁸Bioinformatics and Integrative Genomics PhD Program, Harvard Medical School, Boston, MA, USA. ⁹European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, UK. ¹⁰Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. ¹¹Foundation Medicine Inc., Cambridge, MA, USA. ¹²Wake Forest Comprehensive Cancer Center and Department of Pathology, Wake Forest School of Medicine, Winston-Salem, NC, USA. ¹³Center for Patient Derived Models, Dana-Farber Cancer Institute, Boston, MA, USA. ¹⁴Drug Development Department (DITEP), INSERM U1015, Université Paris Saclay, Gustave Roussy, Villejuif, France. ¹⁵Sorbonne Université, Inserm, CNRS, UMR S 1127, Institut du Cerveau et de la Moelle épinière, ICM, Paris, France. ¹⁶Unité fonctionnelle d'Oncogénétique et Angiogénétique Moléculaire, Département de génétique, Hôpitaux Universitaires La Pitié Salpêtrière - Charles Foix, Paris, France. ¹⁷Department of Diagnostic Radiology, Gustave Roussy, Villejuif, France. ¹⁸IR4M (UMR8081), Université Paris-Sud, Centre National de la Recherche Scientifique, Orsay, France. ¹⁹AP-HP, Université Paris Descartes, Hôpital Cochin, Service d'Anatomie et Cytologie Pathologiques, Paris, France. ²⁰Sorbonne Université, Inserm, CNRS, UMR S 1127, Institut du Cerveau et de la Moelle épinière, ICM, AP-HP, Hôpitaux Universitaires La Pitié Salpêtrière - Charles Foix, Service de Neuropathologie Laboratoire Escourolle, Paris, France. ²¹Sorbonne Université, Inserm, CNRS, UMR S 1127, Institut du Cerveau et de la Moelle épinière, ICM, AP-HP, Hôpitaux Universitaires La Pitié Salpêtrière - Charles Foix, Service de Neurochirurgie, Paris, France. ²²Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, USA. ²³Department of Radiation Oncology, Arthur G. James Hospital/Ohio State Comprehensive Cancer Center, Columbus, OH, USA. ²⁴Department of Neurosurgery, Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA. ²⁵Department of Neurosurgery, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. ²⁶Department of Pathology, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. ²⁷Department of Pediatrics, Harvard Medical School, Boston, MA, USA. ²⁸Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. ²⁹Department of Radiology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. ³⁰Onconeurotek Tumor Bank, Institut du Cerveau et de la Moelle épinière, ICM, Paris, France. ³¹Ludwig Center at Harvard Medical School, Harvard Medical School, Boston, MA, USA. ³²Laboratory of Systems Pharmacology, Harvard Medical School, Boston, MA, USA. ³³Sorbonne Université, Inserm, UMR 938, Centre de Recherche Saint Antoine, Equipe Instabilité des Microsatellites et Cancer, Equipe labellisée par la Ligue Nationale contre le Cancer, Paris, France. ³⁴These authors contributed equally: Mehdi Touat, Yvonne Y. Li. ³⁵These authors jointly supervised this work: Rameen Beroukhim, Pratiti Bhandopadhyay, Franck Bielle, Keith L. Ligon. ✉e-mail: mehdi.touat@gmail.com; rameen_beroukhim@dfci.harvard.edu; pratiti_bhandopadhyay@dfci.harvard.edu; franck.bielle@aphp.fr; keith_ligon@dfci.harvard.edu

Methods

Datasets

For the DFCI-Profile dataset, clinical data and tumour variant calls identified through targeted next-generation sequencing (NGS) panels of 1,628 gliomas sequenced between June 2013 and November 2018 as part of a large institutional prospective profiling program (DFCI-Profile) were included¹⁶ (Extended Data Fig. 1). The distinction between photon and proton radiotherapy was not systematically captured; the vast majority of patients underwent photon radiotherapy. For the MSKCC-IMPACT and FMI datasets, clinical data and tumour variant calls from a total of 545 and 8,121 samples, respectively, that could be assigned to a molecular subgroup (see below) were included^{15,17,32,33}. For pan-cancer analyses in targeted panel sets, clinical data and tumour variant calls from the GENIE project (a repository of genomic data obtained during routine clinical care at international institutions) were downloaded from Synapse (public data, release v6.1)³⁴. For pan-cancer analyses in whole-exome sequencing sets, clinical data and tumour variant calls from 17 hypermutated glioblastomas⁴ and from the pan-cancer TCGA dataset were downloaded from the NCI Genomic Data Commons³⁵. In addition, 247 gliomas collected at one site between 2009 and 2017 were analysed for protein expression of four MMR proteins (MSH2, MSH6, MLH1, and PMS2) using immunohistochemistry. Written informed consent or IRB waiver of consent was obtained from all participants. Patients of the FMI dataset were not consented for release of raw sequencing data. The study, including the consent procedure, was approved by the institutional ethics committees (10-417/11-104/17-000; Western Institutional Review Board (WIRB), Puyallup, WA).

Tumour genotyping and diagnosis

For the majority of samples, genomic testing was ordered by the pathologist or treating physician as part of routine clinical care to identify relevant genomic alterations that could potentially inform diagnosis and treatment decisions. Patients who underwent DFCI-Profile testing signed a clinical consent form, permitting the return of results from clinical sequencing. In total, 1,628 gliomas were sequenced as part of a cohort of 21,992 tumours prospectively profiled between June 2013 and November 2018. Research tumour diagnoses were reviewed and annotated according to histopathology, mutational status of *IDH1* and *IDH2* genes, and whole-arm co-deletion of chromosomes 1p and 19q (1p/19q co-deletion), according to WHO 2016 criteria¹². All samples were assigned to one of four main molecular subgroups: *IDH1/2*-mutant and 1p/19q co-deleted oligodendrogliomas (high- and low-grade), *IDH1/2*-mutant astrocytomas (high- and low-grade), *IDH1/2* wild-type glioblastomas (high-grade only), and *IDH1/2* wild-type gliomas of other histologies (high- and low-grade), the latter including grade I pilocytic astrocytomas, glioneuronal tumours and other unclassifiable gliomas. For simplification, *IDH1/2* wild-type grade III anaplastic astrocytomas and grade IV diffuse intrinsic pontine gliomas were assigned to the group of *IDH1/2* wild-type glioblastomas in all analyses. Samples for which the clinical diagnosis of glioma could not be confirmed (other histology or possible non-tumour sample) and five samples with missing minimal clinical annotation were excluded from all analyses. For the MSKCC-IMPACT and FMI datasets, patients also signed a consent form, and samples were classified using the same procedure. *MGMT* promoter methylation status was determined as part of routine clinical care using chemical (bisulfite) modification of unmethylated, but not methylated, cytosines to uracil and subsequent PCR using primers specific for either methylated or the modified unmethylated DNA in the CpG island of the *MGMT* gene (GenBank accession number AL355531 nt46931-47011).

Targeted panel next-generation sequencing (DFCI-Profile) was performed using the previously validated OncoPanel assay at the Center for Cancer Genome Discovery (Dana-Farber Cancer Institute) for 277 (POPv1), 302 (POPv2), or 447 (POPv3) cancer-associated genes^{16,36}.

In brief, between 50 and 200 ng tumour DNA was prepared as previously described^{16,37}, hybridized to custom RNA bait sets (Agilent SureSelect TM, San Diego, CA) and sequenced using Illumina HiSeq 2500 with 2 × 100 paired-end reads. Sequence reads were aligned to reference sequence b37 edition from the Human Genome Reference Consortium using bwa, and further processed using Picard (version 1.90, <http://broadinstitute.github.io/picard/>) to remove duplicates and Genome Analysis Toolkit (GATK, version 1.6-5-g557da77) to perform localized realignment around indel sites³⁸. Single-nucleotide variants were called using MuTect v1.1.4³⁹, insertions and deletions were called using GATK Indelocator, and variants were annotated using Oncotator⁴⁰. Copy number variants and structural variants were called using the internally developed algorithms RobustCNV⁴¹ and BreakMer⁴² followed by manual review. To filter out potential germline variants, the standard pipeline removes SNPs present at >0.1% in Exome Variant Server, NHLBI GO Exome Sequencing Project (ESP), Seattle, WA (<http://evs.gs.washington.edu/EVS/>, accessed May 30, 2013), present in dbSNP, or present in an in-house panel of normal tissue, but rescues those also present in the COSMIC database⁴³. For this study, variants were further filtered by removing variants present at >0.1% in the gnomAD v2.1.1 database or annotated as benign or likely benign in the ClinVar database^{44,45}. Arm-level copy number changes were generated using an in-house algorithm specific for panel copy number segment files followed by manual expert review. We set a copy number segment mean log₂ ratio threshold at which we could accurately call arm amplification and deletion based on the average observed noise in copy number segments. Chromosome arms were classified as amplified or deleted if more than 70% of the arm was altered. A sample was considered co-deleted if more than 70% of both 1p and 19q were deleted.

Sequencing data from MSKCC-IMPACT were generated at the Memorial Sloan Kettering Cancer Center using a custom targeted panel capture to examine the exons of 341 (IMPACT341) or 398 (IMPACT410) cancer-associated genes as previously described¹⁷. The FMI dataset comprised specimens sequenced as a part of clinical care using a targeted next-generation sequencing assay as previously described (FoundationOne or FoundationOne CDx, Cambridge, MA)^{15,33}. Germline variants without clinical significance were further filtered by applying an algorithm to determine somatic or germline status⁴⁶. Results were analysed for genomic alterations, TMB, MSI and mutational signatures. TMB was assessed by counting all mutations and then excluding germline and known driver mutations^{33,43,47}. The remaining count was divided by the total covered exonic regions^{15,33}. MSI status was determined as previously described⁴⁸. A log-ratio profile for each sample was obtained by normalizing the sequence coverage at all exons and ~3,500 genome-wide SNPs against a process-matched normal control. This profile was corrected for GC-bias, segmented and interpreted using allele frequencies of sequenced SNPs to estimate tumour purity and copy number at each segment. Loss of heterozygosity (LOH) was called if local copy number was 1, or if local copy number was 2 with an estimated tumour minor allele frequency of 0%. To assess 1p/19q co-deletion, we calculated the percentage of each chromosome arm that was monoallelic (under LOH)⁴⁶. A sample was considered 1p/19q co-deleted if both 1p and 19q were >50% monoallelic.

For the DFCI-Profile and FMI datasets, the appropriate cutoffs for hypermutation (17.0 and 8.7 mut/Mb, respectively) were determined by examining the distribution of TMB in all samples and further confirmed using segmented linear regression analysis (Extended Data Fig. 2). For the MSKCC-IMPACT datasets, a threshold previously validated in this dataset was used¹⁷. In all analyses, the homopolymer indel burden was calculated by computing the number of single base insertions or deletions in homopolymer regions of at least 4 bases in length and dividing the count by the total exonic coverage as previously established⁴⁹. Somatic variants were annotated as previously described^{15-17,36,37}. In addition, for the DFCI-Profile and MSKCC-IMPACT datasets, variants in a selected list of glioma- and DNA-repair associated genes (*IDH1*, *IDH2*,

TERT, ATRX, CIC, H3F3A, HIST1H3B, EGFR, PDGFRA, FGFR1, FGFR2, FGFR3, MET, KRAS, NRAS, HRAS, BRAF, NF1, PTPN11, PTEN, PIK3CA, PIK3C2B, PIK3R1, CDKN2A, CDKN2B, CDKN2C, CDK4, CDK6, CCND2, RB1, TP53, MDM2, MDM4, TP53BP1, PPM1D, CHEK1, CHEK2, RAD51, BRCA1, BRCA2, ATM, ATR, MLH1, MLH3, PMS1, PMS2, MSH2, MSH6, EPCAM, SETD2, POLE, POLD1, MUTYH, WRN) were manually reviewed for oncogenicity using several clinical databases for variant annotation (OncoKB, ClinVar, COSMIC, ExAC, and ARUP).

Mutational signature analyses

All variants detected by the sequencing pipeline covered by at least 30× read depth were stringently filtered for germline origin using the gnomAD (population allele frequency greater than 0.1%), and ClinVar (benign or likely benign annotation) databases^{44,45}, as well as manual review of VAF distributions and variants with VAFs consistent with possible germline origin (45–55% or over 95%). The mutational spectrum of variants filtered during these steps was similar to a previously published germline mutation spectrum⁵⁰. Signature analysis was performed for hypermutated samples in a two-step approach starting with the SomaticSignatures package in R for de novo signature extraction within each group⁵¹. To account for the inherent heuristic quality of the NMF approach, the NMF clustering step was repeated 100 times and chosen result was selected based on identifying signatures with the strongest Pearson's correlation coefficients when compared to the 30 well-established COSMIC signatures v2 (https://cancer.sanger.ac.uk/cosmic/signatures_v2)¹⁴ (Extended Data Fig. 5a–c). We then used the DeconstructSigs package in R to estimate the contribution of identified signatures using a regression model⁵². To account for the potential overfitting of a regression approach—owing to either lack of important signatures in the model, or inclusion of uninvolved signatures—we used only the signatures identified by the decomposition approach in step one, supplemented by any strong signature predictions identified through a first pass run of DeconstructSigs with the 30 COSMIC signatures to check for samples that may show strong correlation to an outlier signature. For the FMI dataset, mutational signatures were called as previously described¹⁷. All point mutations were included in the analysis except known oncogenic driver mutations and predicted germline mutations. A sample was deemed to have a dominant signature if a mutational signature had a score of 0.4 or greater.

To assess the ability of this method to detect hypermutation-associated signatures in targeted panel sequencing data, we compared the signature calls of exome-sequenced samples using all variants (previously published DeconstructSigs signature predictions⁵²) versus using only variants that overlapped with the panel-targeted regions. Somatic variant calls for bladder cancer, colon adenocarcinoma, rectal adenocarcinoma, skin cutaneous melanoma, and lung adenocarcinoma (study abbreviations BLCA, COAD, READ, SKCM, LUAD) from the TCGA MC3 dataset were used⁵³ to assess the detection of COSMIC mutational signatures associated with APOBEC (signatures 2 and 13), mismatch repair (signature 6), ultraviolet light (signature 7), POLE (signature 10), and tobacco (signature 4). Variant calls for 17 hypermutated and 12 non-hypermutated glioma exome-sequenced samples were used for assessing temozolomide (signature 11) detection⁴. There were two COAD samples with known POLE exonuclease domain oncogenic mutations and a POLE signature predicted by DeconstructSigs; these were used for assessing POLE signature detection. For a given threshold number of variants (X^1), we considered how many samples had at least X^1 variants, and what percentage of these samples could correctly predict the exome-based signature using panel-restricted variants (with a predicted signature fraction greater than 0.1–0.2). This analysis showed that panel-based signature calls for the APOBEC, mismatch repair, tobacco, and ultraviolet light signatures reached 90% sensitivity with at least 20 somatic variants. Owing to the low number of samples with POLE-associated and temozolomide-associated hypermutation, we did not assess the sensitivity of signature detection at each variant count

threshold; we instead downsampled the number of variants in positive control samples to find the minimum number of variants necessary to reproducibly predict the known signature, which was also determined to be 20 somatic variants (Extended Data Fig. 4).

Enrichment analysis

Mutation enrichment was statistically determined through a permutation test to control for confounders including variable mutability of different genes as well as sample mutation rates, which is of particular importance when assessing enrichment in hypermutated samples. First, we generated a list of every mutation in each of our samples. We calculated the difference in the mutation counts (Δ') between the group of interest and the reference group. We then randomly permuted the mutations 100,000 times, preserving sample and gene mutation counts, and computed the Δ for each gene in each permutation. The P value for a given gene was determined by the fraction of permutations $1-n$ (in our case, $n = 100,000$) for which $\Delta_n \geq \Delta'$. Storey q values were generated using the *qvalue* package in R to adjust for multiple comparisons. The analysis was first performed in the merged DFCI-Profile and MSKCC-IMPACT dataset, and further revalidated in the FMI dataset in an independent analysis.

Single-cell whole-genome sequencing

Frozen glioma samples were mechanically dissociated into pools of single nuclei as previously described⁵⁴, following which single nuclei were isolated by flow cytometry, using a DAPI-based stain. Nuclei were subjected to whole-genome multiple displacement (MDA) amplification (Qiagen, REPLI-g) followed by next-generation sequencing library construction for Illumina Sequencing (Qiagen QIAseq FX DNA library kit). Libraries were sequenced on the Illumina HiSeq platform in paired end mode. Single cells were sequenced to 0.1–1× coverage. Bulk pooled nuclei were sequenced to 60× coverage while matched germline DNA (extracted from blood) was sequenced to 30× coverage.

Reads were aligned to hg38 using *bwa mem*, and variants were jointly called across bulk normal tissue, primary tumour single cells, and recurrent tumour single cells using the GATK best practices pipeline³⁸ without variant quality score recalibration. Somatic mutations in single cells were called if they were monoallelic, had a homozygous reference genotype call but no alternate-allele support in bulk normal tissue, and had at least three supporting reads in a single cell. Germline heterozygous mutations (gHets) were called if they were monoallelic, were found in dbSNP (version 138, <http://www.ncbi.nlm.nih.gov/snp>), and had at least one supporting read and a heterozygous genotype call in bulk normal tissue. To assess sensitivity in each single cell, we computed the fraction of gHets detected with at least three supporting reads, analogous to our procedure for calling somatic mutations. To estimate the total number of somatic mutations present in each cell, we divided the total number of somatic mutations detected by sensitivity. To obtain 95% confidence intervals on the total mutational burden, we modelled the measurement of sensitivity using a beta distribution with Jeffrey's prior, in which the beta parameters (α , β) are equal to the number of detected gHets + 0.5 and the number of undetected gHets + 0.5, respectively. We identified recurrent tumour single cells as hypermutated if their mutational burden was at least 1.5 times the highest mutational burden detected among primary tumour cells.

The method to detect microsatellite mutations was based on read-based phasing^{55,56} and was previously validated using scWGS data from neurons (I.C.-C. et al., manuscript in preparation). First, the human genome was scanned to define a reference set of microsatellite repeats that can be captured using short reads (that is, between 6 and 60 bp) as previously described⁵⁷. Heterozygous SNPs were then detected in the bulk normal sample using the variant caller GATK³⁸. Next, the reads in a given cell mapping to each heterozygous SNP allele detected in the bulk sample and their mates were extracted. If any of the microsatellites in the reference set were covered by these reads, the distribution

Article

of the allelic repeat lengths supported by the data was obtained by collecting the lengths of all intra-read microsatellite repeats mapped to the microsatellite locus under consideration. To discount truncated microsatellite repeats, we required 10-bp flanking sequences (both 5' and 3') of the intra-read microsatellite repeats to be identical to the reference genome. The same procedure was applied to the bulk sample. Finally, the distributions of microsatellite lengths from the single cell and the bulk sample were compared using the Kolmogorov–Smirnov test. The rates of microsatellite instability for each cell were computed as the number of sites mutated divided by the total number of microsatellites for which a call could be made. We applied FDR correction using 0.05 as a threshold for statistical significance, with a minimum of 8 single cell and 10 normal reads required to make a call. All the code is publicly available (<https://github.com/parklab/MSIprofiler>).

Immunohistochemistry

For the revalidation of MMR defects in an independent set, all prospectively collected surgical samples representing consecutive relapses of diffuse glioma following treatment with alkylating agents in adult patients (surgery between 2009 and 2015) were included. An expert neuropathologist reviewed histological samples from the IHC Pitié Salpêtrière cohort (Supplementary Table 2) in order to assess the WHO 2016 integrated diagnosis and to select the tumour areas for immunohistochemistry and for DNA extraction when molecular testing from formalin-fixed paraffin-embedded (FFPE) tissue material was required. Diffuse gliomas harbouring unambiguous positive IDH1(R132H) immunostaining were classified as *IDH1/2*-mutant. *IDH1/2* status was tested by targeted sequencing in all diffuse gliomas harbouring negative or ambiguous IDH1(R132H) immunostaining. *IDH1/2*-mutant diffuse gliomas with loss of ATRX expression in tumour cells were classified as non 1p/19q co-deleted. 1p/19q co-deletion was tested in all *IDH1/2*-mutant diffuse gliomas with maintained ATRX expression. MGMT status was assessed in *IDH1/2* wild-type gliomas. FFPE sections (3 µm thick) were deparaffinized and immunolabelled with a Ventana Benchmark XT stainer (Roche, Basel, Switzerland). The secondary antibodies were coupled to peroxidase with diaminobenzidine as brown chromogen. For immunohistochemistry performed at Pitié-Salpêtrière (PSL) Hospital, the following antibodies were used: mouse monoclonal anti-ATRX (Bio SB, clone BSB-108, BSB3296, 1:100), mouse monoclonal anti-IDH1(R132H) (Dianova, clone H09, DIA-H09, 1:100), rabbit monoclonal anti-CD3 (Roche, clone 2GV6, 790-4341, prediluted), rabbit polyclonal anti-IBA1 (Wako, W1W019-19741, 1:500), mouse monoclonal anti-MLH1 (Roche, clone M1, 790-4535, prediluted), mouse monoclonal anti-MSH2 (Roche, clone G219-1129, 760-4265, prediluted), mouse monoclonal anti-MSH6 (Roche, clone 44, 760-4455, prediluted), rabbit monoclonal anti-PMS2 (Roche, clone EPR3947, 760-4531, prediluted). For immunohistochemistry performed at BWH, the following antibodies were used: mouse monoclonal anti-MLH1 (Leica, clone ESO5, MLH1-L-CE, 1:75), mouse monoclonal anti-MSH2 (Merck Millipore, clone Ab-2-FE11, NA27, 1:200), mouse monoclonal anti-MSH6 (Leica, clone PU29, MSH6-L-CE, 1:50), mouse monoclonal anti-PMS2 (Cell Marque, MRQ-28, 288M-14-ASR, 1:100). An expert neuropathologist blinded to the molecular status of MMR deficiency analysed the immunostaining. If loss of expression of one or several MMR proteins was observed in tumour cells, this result was confirmed in an independent laboratory by a second expert pathologist with separate stainer and reagents: FFPE sections were immunolabelled with a BOND stainer (Leica, Wetzlar, Germany). Primary antibodies were as follows: mouse monoclonal anti-MLH1 (clone G168-728, BD Pharmingen), mouse monoclonal anti-MSH2 (clone 25D12, Diagnostic BioSystems), mouse monoclonal anti-MSH6 (clone 44, Diagnostic BioSystems), mouse monoclonal anti-PMS2 (clone A16-4, BD Pharmingen). The loss of expression of MMR proteins was defined as the total absence of nuclear labelling in tumour cells associated with a maintained expression in normal cells (as a positive internal control in the same tissue area). The density of

the immune infiltrate was studied after immunolabelling of T lymphocytes by CD3 and of macrophage/microglial cells by IBA1. The number of immunopositive cells was quantified by visual counting in the three areas (one square millimetre) of tumour tissue containing the highest density of immunopositive cells and a mean density was calculated.

Patient-derived cell lines

All PDCLs with a name starting with BT were established from tumours resected at Brigham and Women's Hospital and Boston Children's Hospital (Boston, MA) and were maintained in neurosphere growth conditions using the NeuroCult NS-A Proliferation Kit (StemCell Technologies) supplemented with 0.0002% heparin (StemCell Technologies), EGF (20 ng/ml), and FGF (10 ng/ml; Miltenyi) in a humidified atmosphere of 5% CO₂ at 37 °C. The N16-1162 PDCL was established by the GlioTex team (Glioblastoma and Experimental Therapeutics) at the Institut du Cerveau et de la Moëlle épinière (ICM) laboratory and maintained as described above. SU-DIPG-XIII (DIPG13) cells were provided by Dr. Michelle Monje at Stanford University and were maintained in neurosphere growth conditions in a humidified atmosphere of 5% CO₂ at 37 °C in tumour stem medium (TSM) consisting of Dulbecco's modified Eagle's medium: nutrient mixture F12 (DMEM/F12), neurobasal-A medium, HEPES buffer solution 1 M, sodium pyruvate solution 100 nM, non-essential amino acids solution 10 mM, Glutamax-1 supplement and antibiotic-antimycotic solution (Thermo Fisher). The medium was supplemented with B-27 supplement minus vitamin A, (Thermo Fisher), 20 ng/ml human-EGF (Miltenyi), 20 ng/ml human-FGF-basic (Miltenyi), 20 ng/ml human-PDGF-AA, 20 ng/ml human-PDGF-BB (Shenandoah Biotech) and 2 µg/ml heparin solution (0.2%, Stem Cell Technologies). The identity of all cell lines established was confirmed by short tandem repeat assay or sequencing. All cell lines were tested for the absence of mycoplasma. Cell lines, xenografts, and model data available from the DFCI Center for Patient Derived Models.

Viability assays

For short-term viability assays, cells were plated in 96-well plates and treated the following day with temozolomide (Selleckchem) or CCNU (Selleckchem) for 7–9 days incubation. Fresh medium was added after four days of incubation. Cell viability was assessed using the luminescent CellTiter-Glo reagent (Promega) according to the manufacturer's protocol. Luminescence was measured using the Modulus Microplate Reader (Promega). The surviving fraction (SF) for each [X] concentration was calculated as SF = mean viability in treated sample at concentration [X]/mean viability of untreated samples (vehicle). Dose–response curves and IC₅₀ were generated using Prism 8 (GraphPad Software, San Diego, USA) after log transformation of the concentrations. Curves were extrapolated using nonlinear regression with four-parameter logistic regression fitting on triplicates from survival fractions of three independent replicates, following the model: $y = \text{Bottom} + (\text{Top} - \text{Bottom}) / (1 + 10^{([\log \text{IC}_{50} - x] \times \text{HillSlope}))$.

Generation of Isogenic MMR-deficient cell lines

Oligos of the form 5'-CACCG[N20] (where [N20] is the 20-nucleotide target sequence; sgFDP, GAGCTGGACGGCGACGTAAA; sgMSH2, ATTCTGTCTTATCCATGAC; sgMSH6, TTATTGGAGTCAGTGAACGT; sgMLH1, ACTACCCAATGCCTCAACCG; sgPMS2, TCATGCAGCAGCGAGTATG) and 5'-AAAC[rc20]C (where [rc20] is the reverse complement of [N20]) were purchased from Integrated DNA Technologies (IDT). For DIPG13 cells, oligos containing the sgRNA target sequence were annealed with their respective reverse complement and cloned into the lentiCRISPR all-in-one sgRNA/Cas9-delivery lentiviral expression vector (pXPR_BRD001; now available as pXPR_BRD023 lentiCRISPRv2) from the Broad Institute Genetics Perturbation Platform (GPP). For BT145 cells, oligos containing the sgRNA target sequence were annealed with their respective reverse complement and cloned into the pXPR_BRD051

CRISPRko all-in-one sgRNA/Cas9-delivery lentiviral expression vector (available from the Broad Institute Genetics Perturbation Platform, GPP). Successful cloning of each sgRNA target sequence was confirmed via Sanger Sequencing. To generate lentivirus from these vectors, HEK293T cells were transfected with 10 µg of each expression plasmid with packaging plasmids encoding PSPAX2 and VSVG using lipofectamine. Lentivirus-containing supernatant was collected 48 and 72 h after transfection. DIPG13 and BT145 cells were seeded in a 12-well plate at $1-3 \times 10^6$ cells/well in 3 ml medium and spin-infected (2,000 rpm for 2 h at 30 °C with no polybrene) with pLX311-Cas9 (DIPG13) or pXPR_BRD051 (BT145) lentiviral vectors and selected with blasticidin (10 µg/ml, DIPG13) or hygromycin (300 µg/ml, BT145) to generate Cas9-expressing or knockout cells. DIPG13-Cas9 cells underwent a subsequent lentiviral spin-infection with the lentiCRISPR sgGFP, sgMSH2, or sgMSH6 vectors described above. Puromycin selection (0.4 µg/ml for DIPG13 cells) commenced 48 h post-infection.

Chronic treatment and sequencing of isogenic MMR-deficient cell lines

DIPG13-sgGFP, -sgMSH2, and -sgMSH6 cells were seeded at 8×10^5 cells/well in 4 ml medium in a 6-well ULA plate. Each line was grown for 3 months under 3 conditions: no treatment, temozolomide (100 µM, Selleckchem), or DMSO vehicle. Cells were grown under these conditions in the absence of both blasticidin and puromycin. Cells were re-dosed with temozolomide or DMSO every 3–5 days, splitting over-confluent cells 1:2 or 1:4 as needed. After 3 months, genomic DNA was extracted using the QIAmp DNA Mini Kit. DNA was subjected to whole-exome Illumina sequencing. Reads were aligned to the Human Genome Reference Consortium build 38 (GRCh38). WES data were analysed using the Getz Lab CGA whole-exome sequencing characterization pipeline (https://docs.google.com/document/d/1VO2kX_gfUd0x3mBS9NjLUWGZu794WbTepBel3cBg08/edit#heading=h.yby87l2ztbcj) developed at the Broad Institute which uses the following tools for quality control, calling, filtering and annotation of somatic mutations and copy number variation: PicardTools (<http://broadinstitute.github.io/picard/>) ContEst⁵⁸, MuTect1³⁹, Strelka⁵⁹, Orientation Bias Filter⁶⁰, DeTiN⁶¹, AllelicCapSeg⁶², MAFFPoNFilter⁶³, RealignerFilter, ABSOLUTE⁶⁴, GATK³⁸, Variant Effect Predictor⁶⁵, and Oncotator⁴⁰.

Subcutaneous xenografts and drug treatment

BT145 cells (2×10^6) were resuspended in equal parts Hank's buffered salt solution (Life Technologies) and Matrigel (BD Biosciences) and then injected into both flanks of eight-week-old NU/NU male mice (Charles River). Tumour-bearing mice ($n = 8$) were randomly assigned to the treatment or vehicle arm when tumours measured a volume of 100 mm³. Animals received 12 mg/kg/day temozolomide or vehicle (Ora-Plus oral suspension solution, Perrigo, Balcatta, Australia) by oral gavage for 5 consecutive days per 28-day cycle. An additional 4 weeks resting period without treatment was observed before the second cycle. Tumour volumes were calculated using the formula: $0.5 \times \text{length} \times \text{width}^2$. Body weights were monitored twice weekly. The investigators were not blinded to allocation during experiments and outcome assessment. Mice were euthanized when they showed signs of tumour-related illness or before reaching the maximum tumour burden. Tumours were subsequently removed, and a subset were submitted to Oncopanel sequencing for analyses of exonic mutations (POPv3, 447 genes) and mutational signature as defined above. To separate human and mouse sequenced reads in the DNA sequencing data generated for the PDX models, the 'raw' data were mapped to both the hg19 human and mm10 mouse reference genomes using BWA-MEM-0.7.17. The output of the alignment was name sorted by Samtools-1.7. We then used the software package Disambiguate (ngs_disambiguate-1.0) to assign each read to the human or mouse genome and to produce final alignment files in BAM format. Final hg19 BAM files were coordinate sorted by Samtools-1.7. Duplicate reads were marked and removed from the

BAM files using Picard-2.0.1. GATK4.1.0.0 base recalibration was performed using BaseRecalibration and Applying Recalibration followed by CollectF1R2Counts and LearnReadOrientationModel to create a model for read orientation bias. Variant calling was performed using GATK-4.1.0.0/Mutect2 pipeline with the default parameters and filters except for the following modifications: (i) 'af-of-alleles-not-in-resource' was set to 0; (ii) 'MateOnSameContigOrNoMappedMateReadFilter' was disabled; (iii) the output of Step8 was used for fitting the read orientation model; and (iv) a germline resource from the gnomAD database was included (https://console.cloud.google.com/storage/browser/_details/gatk-best-practices/somatic-b37/af-only-gnomad.raw.sites.vcf). The capture targets intervals used for Mutect2 were POPv3. The generated variant calls were further filtered using the FilterMutectCalls module of GATK4.1.0.0 and the final output in VCF format was annotated with Ensemble Variant Effect Predictor (ensembl-vep-96.0) using vcf-2maf-1.6.16. The calls were additionally annotated with the OncoKB dataset using oncoKB-annotator and sorted as MAF files.

All in vivo studies were performed in accordance with Dana-Farber Cancer Institute animal facility regulations and policies under protocol number 09-016.

Immunoblotting

Proteins were extracted in lysis RIPA buffer (50 mM Tris, 150 mM NaCl, 5 mM EDTA, 0.5% sodium deoxycholic acid, 0.5% NP-40, 0.1% SDS) supplemented with protease inhibitor cocktail (Roche Molecular). Proteins were quantified using the PierceBCA Protein Assay Kit, according to the manufacturer's protocol. Samples were then prepared with 1× NuPAGE (Invitrogen) LDS sample buffer, and NuPAGE (Invitrogen) sample reducing agent followed by heating to 95 °C for 5 min. The samples were then loaded onto NuPAGE 4–12% Bis-Tris Gel (Invitrogen) with NuPAGE MOPS SDS (Invitrogen) buffer and run through electrophoresis. The transfer onto membrane was then done at 40 V overnight at 4 °C in NuPAGE transfer buffer (Invitrogen) with 10% methanol. Membranes were blocked with 5% skim milk in TBST for 1 h, then incubated with the following primary antibodies added to 5% BSA and incubated overnight at 4 °C on a shaker: mouse monoclonal anti-MGMT (Millipore, MT3.1, MAB16200, 1:500), mouse monoclonal anti-MSH2 (Calbiochem, FE11, NA27, 1:1,000), mouse monoclonal anti-MSH6 (Biosciences, 44, 610918, 1:500), mouse monoclonal anti-MLH1 (Cell Signaling, 4C9C7, 3515, 1:500), mouse monoclonal anti-PMS2 (BD Biosciences, A16-4, 556415, 1:1,000), mouse monoclonal anti-beta-actin (Sigma, AC-74, A2228, 1:10,000). After several cycles of washing and incubation with secondary goat anti-mouse antibody (Invitrogen 31430, 1:10,000), membranes were imaged by chemiluminescence using the Biorad ChemidocTM MP imaging system.

Microsatellite instability analysis

PCR amplification of the five mononucleotide markers (BAT25, BAT26, NR21, NR24, MONO27) was performed with the MSI Analysis System kit (Version 1.2, Promega). PCR products were analysed by an electrophoretic separation on the polymer POP7 50cm in an Applied Biosystems 3130XL sequencer and using Genemapper Software 5.

Outcome of patients treated with PD-1 blockade

For comparison of PFS and OS in patients treated with PD-1 pathway blockade according to TMB and MMR statuses, we retrospectively identified patients with glioma who had been treated with PD-1 blockade (alone or in combination with bevacizumab) for recurrent disease at our institutions. Patients for whom sequencing was not performed at the time of recurrence were excluded. Magnetic resonance imaging (MRI) tumour assessments were reviewed using the Response Assessment in Neuro-Oncology (RANO) criteria by three independent reviewers (M.J.L.-F., S.A., and R.Y.H.) who were blinded to the groups. PFS and OS duration were calculated from cycle 1 day 1 of PD-1 blockade therapy.

Article

Statistical analyses

Data were summarized as frequencies and proportions for categorical variables and as median and range for continuous variables. Continuous variables were compared using Mann–Whitney or Kruskal–Wallis tests; categorical variables were compared using Fisher's exact or Chi-squared tests. Survival and PFS were estimated using the Kaplan–Meier method, and differences in survival or PFS between groups were evaluated by the log-rank test. Survival for subjects who were alive or lost to follow-up at the time of last contact on or before data cut-off was censored at the date of the last contact. Patient matching in a k -to- k fashion was conducted using coarsened exact matching according to diagnosis, primary versus recurrent status, and prior treatments. For evaluation of response to PD-1 blockade, patients with glioma from the DFCI-Profile cohort who were treated with anti-PD(L)-1 antibodies or other treatments (total $n = 210$) as part of their management were included in the analysis. For multivariable analysis, Cox proportional hazard regression was used to investigate the variables that affect survival. P values were considered statistically significant when <0.05 . Statistical analyses were performed using STATA (v14.2, StataCorp LLC, College Station, USA), Prism 8 (GraphPad Software, San Diego, USA), and MedCalc Statistical Software, version 19.0.3 (MedCalc Software bvba, Ostend, Belgium). For enrichment analyses, mutated genes were considered significant when $Q < 0.01$. Where applicable, the means of population averages from multiple independent experiments (\pm s.d. or s.e.m.) are indicated. No statistical methods were used to pre-determine sample size.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Clinical and sequencing data from 1,495 samples from the DFCI-Profile and 545 samples from the MSKCC-IMPACT datasets are publicly available (GENIE v.6.1: <https://genie.cbioportal.org> or <https://www.synapse.org/>). All data for samples from the GENIE v.6.1 and TCGA pan-cancer datasets are publicly available. Data for samples from the FMI dataset are not publicly available, but de-identified, aggregated data can be accessed on request. dbGaP Study Accession: phs001967.v1.p1. All other data are available on request.

Code Availability

The code for the detection of microsatellite mutations in single-cell DNA sequencing is publicly available (<https://github.com/parklab/MSIprofiler>).

32. Jonsson, P. et al. Genomic correlates of disease progression and treatment response in prospectively characterized gliomas. *Clin. Cancer Res.* **25**, 5537–5547 (2019).
33. Chalmers, Z. R. et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome Med.* **9**, 34 (2017).
34. Consortium, A. P. G.; AACR Project GENIE Consortium. AACR Project GENIE: powering precision medicine through an international consortium. *Cancer Discov.* **7**, 818–831 (2017).
35. Cancer Genome Atlas Research Network et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45**, 1113–1120 (2013).
36. Garcia, E. P. et al. Validation of OncoPanel: a targeted next-generation sequencing assay for the detection of somatic variants in cancer. *Arch. Pathol. Lab. Med.* **141**, 751–758 (2017).
37. Ramkissoon, S. H. et al. Clinical targeted exome-based sequencing in combination with genome-wide copy number profiling: precision medicine analysis of 203 pediatric brain tumors. *Neuro-oncol.* **19**, 986–996 (2017).
38. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
39. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
40. Ramos, A. H. et al. OncoPrint: cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).
41. Bi, W. L. et al. Clinical identification of oncogenic drivers and copy-number alterations in pituitary tumors. *Endocrinology* **158**, 2284–2291 (2016).
42. Abo, R. P. et al. BreakMer: detection of structural variation in targeted massively parallel sequencing data using kmers. *Nucleic Acids Res.* **43**, e19 (2015).
43. Tate, J. G. et al. COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
44. Karczewski, K. J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. Preprint at <https://www.biorxiv.org/content/10.1101/531210v2> (2019).
45. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
46. Sun, J. X. et al. A computational approach to distinguish somatic vs. germline origin of genomic alterations from deep sequencing of cancer specimens without a matched normal. *PLOS Comput. Biol.* **14**, e1005965 (2018).
47. Garofalo, A. et al. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome Med.* **8**, 79 (2016).
48. Trabucco, S. E. et al. A novel next-generation sequencing approach to detecting microsatellite instability and pan-tumor characterization of 1000 microsatellite instability-high cases in 67,000 patient samples. *J. Mol. Diagn.* **21**, 1053–1066 (2019).
49. Papke, D. J. Jr et al. Validation of a targeted next-generation sequencing approach to detect mismatch repair deficiency in colorectal adenocarcinoma. *Mod. Pathol.* **31**, 1882–1890 (2018).
50. Rahbari, R. et al. Timing, rates and spectra of human germline mutation. *Nat. Genet.* **48**, 126–133 (2016).
51. Gehring, J. S., Fischer, B., Lawrence, M. & Huber, W. SomaticSignatures: inferring mutational signatures from single-nucleotide variants. *Bioinformatics* **31**, 3673–3675 (2015).
52. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
53. Ellrott, K. et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell Syst.* **6**, 271–281.e277 (2018).
54. Francis, J. M. et al. EGFR variant heterogeneity in glioblastoma resolved through single-nucleus sequencing. *Cancer Discov.* **4**, 956–971 (2014).
55. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
56. Bohrsen, C. L. et al. Linked-read analysis identifies mutations in single-cell DNA-sequencing data. *Nat. Genet.* **51**, 749–754 (2019).
57. Cortes-Ciriano, I., Lee, S., Park, W. Y., Kim, T. M. & Park, P. J. A molecular portrait of microsatellite instability across multiple cancers. *Nat. Commun.* **8**, 15180 (2017).
58. Cibulskis, K. et al. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* **27**, 2601–2602 (2011).
59. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* **28**, 1811–1817 (2012).
60. Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* **41**, e67 (2013).
61. Taylor-Weiner, A. et al. DeTiN: overcoming tumor-in-normal contamination. *Nat. Methods* **15**, 531–534 (2018).
62. Landau, D. A. et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714–726 (2013).
63. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
64. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
65. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

Acknowledgements We thank the patients and families who took part in the study, as well as the staff, research coordinators and investigators at each participating institution. M.T. is supported by Fondation pour la Recherche Médicale (FDM 41635), Fondation Monahan, The Arthur Sachs Foundation and The Philippe Foundation. C.L.B. was funded by a Bioinformatics and Integrative Genomics training grant from NHGRI (T32HG002295). S.S. is supported by the Ludwig Center at Harvard. M.S. is supported by Institut National du Cancer (INCa), the Ligue Nationale contre le Cancer (Equipe Labelisée), and Investissements d'avenir. R.B. is supported by NIH R01 CA188228, R01 CA215489, and R01 CA219943, The Dana-Farber/Novartis Drug Discovery Program, The Gray Matters Brain Cancer Foundation, Ian's Friends Foundation, The Bridge Project of MIT and Dana-Farber/Harvard Cancer Center, The Pediatric Brain Tumor Foundation, the Fund for Innovation in Cancer Informatics, and The Sontag Foundation. P.B. is supported by NIH K99 CA201592, R00CA201592-03, the Dana-Farber Cancer Institute and Novartis Institute of Biomedical Research Drug Discovery and Translational Research Program, the Pediatric Brain Tumor Foundation and the St Baldrick's Foundation. F. Bielle is supported by Fondation ARC pour la recherche sur le cancer (PJA 20151203562), INCa, a grant Emergence (Sorbonne Université) and ARCT (Association pour la recherche sur les tumeurs cérébrales). K.L.L. is supported by R01CA188288, P01 CA163205, P50 CA165962, Pediatric Brain Tumor Foundation, and the Ivy Foundation. This work was in part supported by a the SIRIC CURAMUS, which is funded by INCa, the French Ministry of Solidarity and Health and Inserm (INCA-DGOS-Inserm_12560). We acknowledge K. Bryan, S. Valentin, B. Bonneau, A. Matos and I. Deltrait for preparation and processing of samples; W. Pisano and S. Block for help in data collection; E. F. Cohen for mouse xenograft sequencing analyses; D. X. Jin and J. Moore for assistance with FMI dataset creation and curation; the members of the BWH Center for Advanced Molecular Diagnostics; Y. Marie, J. Gueguan and the ICM Genotyping and Sequencing Core Facility (IGENSEQ) for sharing expertise related to analysis of copy array and sequencing data; C. Perry and the DFCI Oncology Data Retrieval System (OncDRS) for the aggregation, management, and delivery of the operational research data used in this project; the American Association for Cancer Research and its financial and material support in the development of the AACR Project GENIE registry, and members of the consortium for their commitment to data sharing; the cBioPortal for Cancer Genomics (<https://www.cbioportal.org>) and the Memorial Sloan Kettering Cancer Center for data sharing of the MSKCC-IMPACT

dataset. We greatly appreciate feedback and support from M. L. Meyerson regarding bioinformatics and genomics analysis, I. K. Mellingshoff and T. J. Kaley for scientific advice, V. Rendo for scientific review of the manuscript, and M. Monje for providing the DIPG13 parental cell line. The content is solely the responsibility of the authors.

Author contributions M.T., R.B., P.B., F. Bielle and K.L.L. designed the study. M.T., Y.Y.L., R.B., P.B., F. Bielle and K.L.L. wrote the initial draft, with input from all authors. Y.Y.L. validated mutational signature analyses using TCGA data. M.T., Y.Y.L., L.F.S., R.S., D. Pavliak and L.A.A. performed TMB and mutational signature analyses of the DFCI-Profile, MSKCC-IMPACT and FMI datasets, and integrated TMB, signature and clinical data. L.F.S. developed the code for permutation tests. M.T., Y.Y.L., L.F.S. and R.S. performed and analysed the permutation tests. M.T., A.N.B., K.P., C. Bellamy, N.C., J.B., K.Q., P.H., S.M., L.T., R.B., P.B. and K.L.L. performed in vitro experiments in native and engineered models and analysed experimental data. M.T., K.P., J.B. and K.L.L. performed in vivo experiments in native models and analysed experimental data. M.T., F. Beuvon, K.M., S. Alexandrescu, D.M.M., S.S., F. Bielle, and K.L.L. reviewed histological and immunohistochemistry data on human samples. M.T. and J.B.I. performed survival analyses. M.T., Y.Y.L., C.L.B., I.C.-C., P.J.P., R.B., P.B. and K.L.L. performed single-cell sequencing experiments and analysed data. C.L.B., I.C.-C. and P.J.P. developed computational tools for the analysis of single-cell data. M.T., Y.Y.L., L.F.S., C.L.B., I.C.-C., S.H.R., F.D., A.S., R.S., D. Pavliak, L.A.A., E.G., G.M.F., F.C., A.D., A. Cherniack, P.J.P., R.B., P.B. and K.L.L. reviewed and analysed the bulk sequencing genomic data. M.T., J.B.I., C. Birzu, J.E.G., M.J.L.-F., R.J., N.Y., C. Baldini, E.G., S. Ammari, F. Beuvon, K.M., A.A., C.D., C.H., F.L.-D., D. Psimaras, E.Q.L., L.N., J.R.M.-F., A. Carpentier, P.C., L.C., B.M., J.S.B.-S., A. Chakravarti, W.L.B., E. A. Chiocca, K.P.F., S. Alexandrescu, S.C., D.H.-K., T.T.B., B.M.A., R.Y.H., A.H.L., F.C., J.-Y.D., K.H.-X., D.M.M., S.S., M.S., P.Y.W., D.A.R., A.M., A.I., R.B., P.B., F. Bielle, and K.L.L. abstracted and reviewed clinical and treatment response data. Y.Y.L., L.F.S., C.L.B., I.C.-C., R.S., L.A.A., G.M.F., A. Cherniack, and R.B. created bioinformatics tools and systems to support data analysis. R.B., P.B., F. Bielle, and K.L.L. acquired funding and supervised the study. All authors participated in data analysis and approved the final manuscript.

Competing interests M.T. reports consulting or advisory role for Agios Pharmaceutical, Integragen, and Taiho Oncology, outside the submitted work; travel, accommodations, expenses from Merck Sharp & Dome, outside the submitted work. Y.Y.L. reports equity from g. Root Biomedical. S.H.R., R.S., D. Pavliak, L.A.A., G.M.F. and B.M.A. report employment with Foundation Medicine and stock interests from Roche. K.M. reports advisory board honoraria from Bristol-Meyers Squibb, outside the submitted work. F.L.-D. reports fees from Pharmtrace, outside the submitted work. E.Q.L. reports consulting or advisory role for Eli Lilly; royalties from UpToDate; honoraria from Prime Oncology. L.N. reports consulting or advisory role for Bristol-Meyers Squibb, outside the submitted work. T.T.B. reports honoraria from Champions Oncology, UpToDate, Immedex, NXDC, Merck, GenomiCare Biotechnology; consulting or advisory role for Merck, GenomiCare Biotechnology, NXDC, Amgen; travel, accommodations, expenses from Merck, Roche, Genentech, GenomiCare Biotechnology. A.H.L. reports leadership from Travera (I); stock and other ownership interests from Travera (I); consulting or advisory role for Travera (I). K.H.-X. reports advisory board honoraria from Bristol-Meyers Squibb, outside the submitted work. S.S. reports personal fees from Rarecyte, outside the submitted work. P.Y.W. reports honoraria from Merck; consulting or advisory role for AbbVie, Agios Pharmaceuticals, AstraZeneca, Blue Earth Diagnostics, Eli Lilly, Genentech, Roche,

Immunomic Therapeutics, Kadmon Corporation, KIYATEC, Puma Biotechnology, Vascular Biogenics, Taiho Pharmaceutical, Deciphera Pharmaceuticals, VBI Vaccines; speakers' bureau from Merck, prime Oncology; research funding from Agios Pharmaceuticals (Inst), AstraZeneca (Inst), BeiGene (Inst), Eli Lilly (Inst), Roche (Inst), Genentech (Inst), Karyopharm Therapeutics (Inst), Kazia Therapeutics (Inst), MediciNova (Inst), Novartis (Inst), Oncocotics (Inst), Sanofi (Inst), Aventis (Inst), VBI Vaccines (Inst); travel, accommodations, expenses from Merck. D.A.R. reports honoraria from AbbVie, Cavion, Genentech, Roche, Merck, Midatech Pharma, Momenta Pharmaceuticals, Novartis, Novocure, Regeneron Pharmaceuticals, Stemline Therapeutics, Celldex, OXiGENE, Monteris Medical, Bristol-Myers Squibb, Juno Therapeutics, Inovio Pharmaceuticals, Oncorus, Agenus, EMD Serono, Merck, Merck KGaA, Taiho Pharmaceutical, Advantagene; consulting or advisory role for Cavion, Genentech, Roche, Merck, Momenta Pharmaceuticals, Novartis, Novocure, Regeneron Pharmaceuticals, Stemline Therapeutics, Bristol-Myers Squibb, Inovio Pharmaceuticals, Juno Therapeutics, Celldex, OXiGENE, Monteris Medical, Midatech Pharma, Oncorus, AbbVie, Agenus, EMD Serono, Merck, Merck KGaA, Taiho Pharmaceutical; research funding from Celldex (Inst), Incyte (Inst), Midatech Pharma (Inst), Tragara Pharmaceuticals (Inst), Inovio Pharmaceuticals (Inst), Agenus (Inst), EMD Serono (Inst), Acerta Pharma (Inst), Omnixox. A.I. reports grants and other from Carthera (September 2019); research grants from Transgene; grants from Sanofi, and Air Liquide; and travel funding from Leo Pharma, outside the submitted work. R.B. reports consulting or advisory role for Novartis, Merck (I), Gilead Sciences (I), Viiv Healthcare (I); research funding from Novartis; patents, royalties, other intellectual property—Prognostic Marker for Endometrial Carcinoma (US patent application 13/911456, filed June 6, 2013), SF3B1 Suppression as a Therapy for Tumors Harboring SF3B1 Copy Loss (international application No. WO/2017/177191, PCT/US2017/026693, filed July 4, 2017), Compositions and Methods for Screening Pediatric Gliomas and Methods of Treatment Thereof (international application No. WO/2017/132574, PCT/US2017/015448, filed 1/27/2017). P.B. reports research grants from the Novartis Institute of Biomedical Research; patents, royalties, other intellectual property—Compositions and Methods for Screening Pediatric Gliomas and Methods of Treatment Thereof (international application No. WO/2017/132574, PCT/US2017/015448, filed 1/27/2017). F. Bielle reports employment from Celgene (I); stocks from Crossject (I); research grants from Sanofi and Abbvie; travel, accommodations, expenses from Bristol-Myers Squibb for travel expenses, outside the submitted work. K.L.L. reports grants and personal fees from BMS, grants from Amgen, personal fees and other from Travera LLC, personal fees from InteraGen, personal fees from Rarecyte, grants from Tragara, grants from Lilly, grants from Deciphera, grants from X4, all outside the submitted work; and has patent US20160032359A1 pending. Inst. denotes institutional funding; I denotes a competing interest involving a first degree relative of the author. The other authors report no competing interests.

Additional information

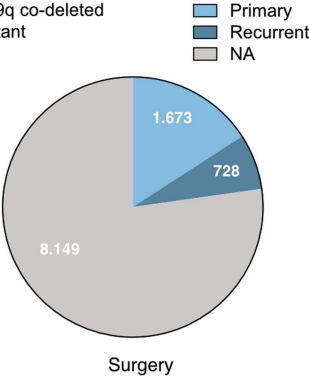
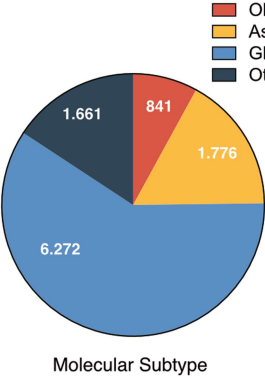
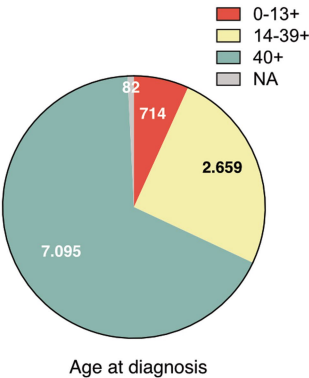
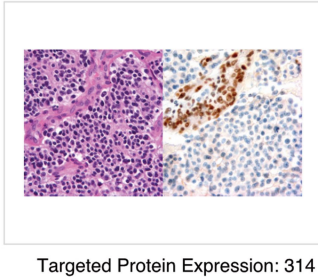
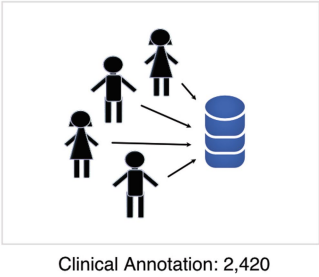
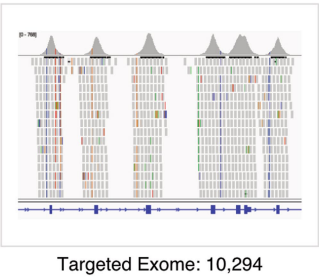
Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2209-9>.

Correspondence and requests for materials should be addressed to M.T., R.B., P.B., F.B. or K.L.L.

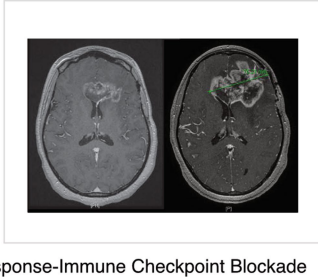
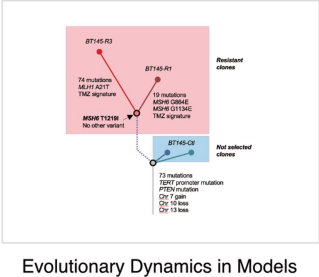
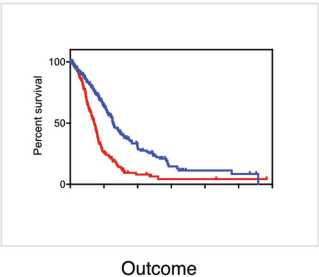
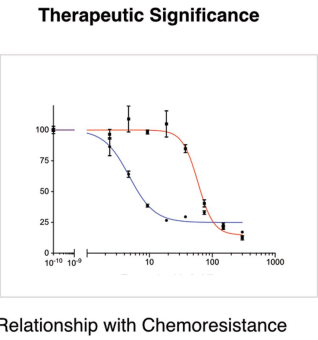
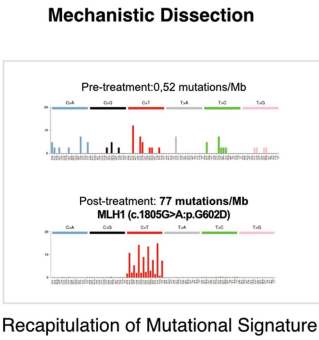
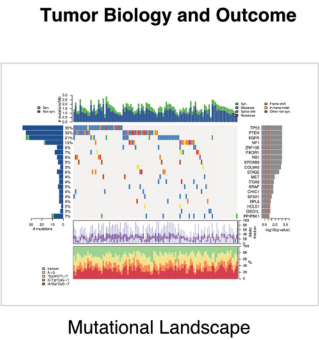
Peer review information Nature thanks Stefan Pfister and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

a

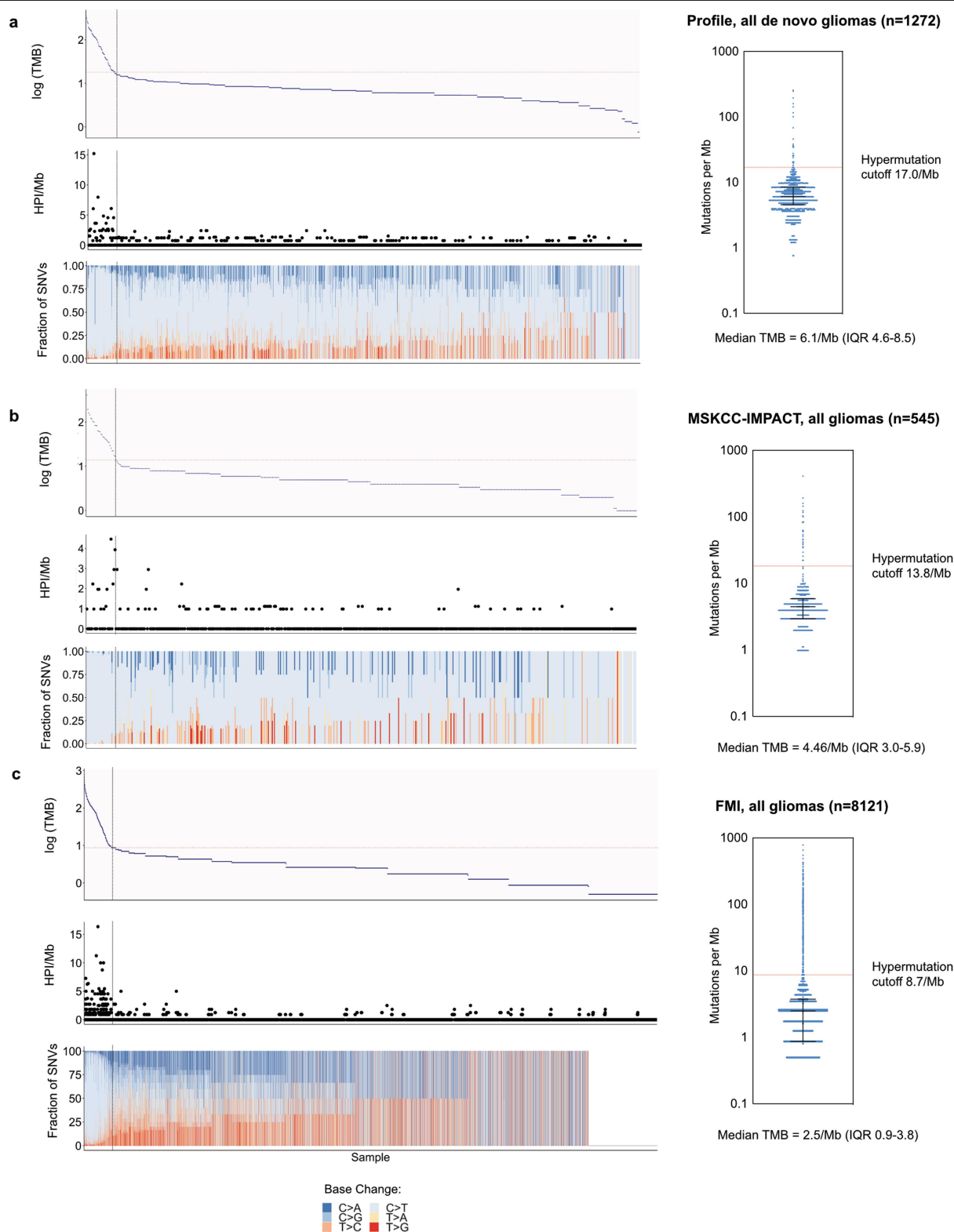


b



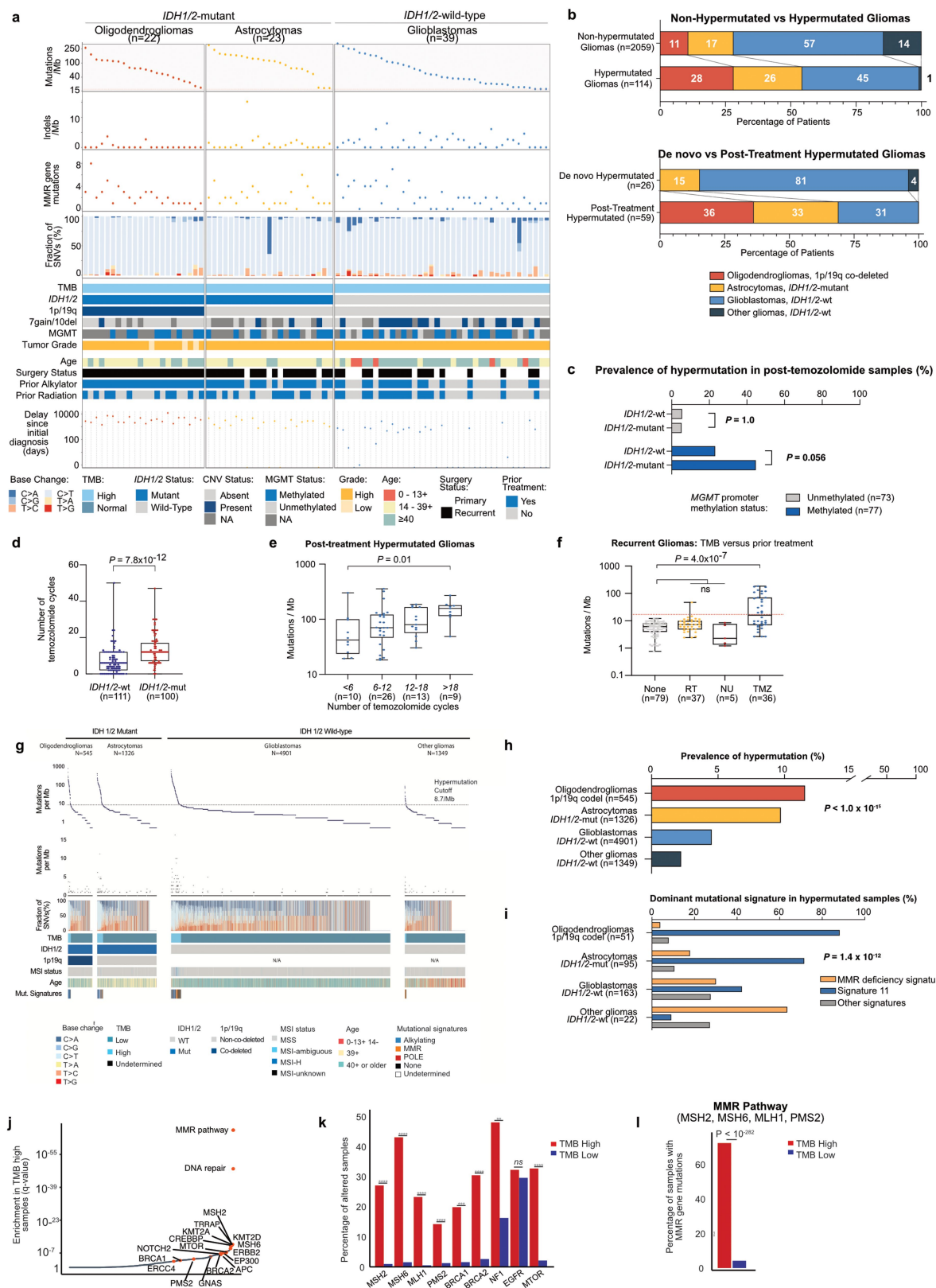
Extended Data Fig. 1 | Overview of the clinical characteristics of the patients in the study and analyses performed. **a**, Clinical datasets analysed and main demographics including age, histomolecular subtype and disease stage. 1,628 glioma samples from adult and paediatric patients were sequenced as part of a large institutional prospective sequencing program of consented patients (DFCI-Profile) and subsequently clinically annotated. We identified

545 and 8,121 gliomas with sequencing from the MSKCC-IMPACT and FMI datasets, respectively, and used them as a replication set (total set of 10,294 sequenced samples). In addition, 314 tumours—including 247 consecutive recurrent gliomas—were analysed for protein expression of four MMR proteins (MSH2, MSH6, MLH1, and PMS2) using immunohistochemistry. **b**, Analyses performed and key clinical questions that were addressed in the study.



Extended Data Fig. 2 | Distributions of TMB, homopolymer indels, and SNV mutation spectra in the datasets used. a, DFCI-Profile (de novo gliomas only); **b**, MSKCC-IMPACT; **c**, FMI (total $n = 9,938$). After examining the distribution of TMB in each dataset for breakpoints, thresholds for hypermutation were further confirmed using segmented linear regression analysis (analysis restricted to de novo gliomas for DFCI-Profile). This method showed the presence of a breakpoint at 17.0 and 8.7 mutations per Mb for the DFCI-Profile and FMI datasets, respectively. For the MSKCC-IMPACT dataset, the cutoff used

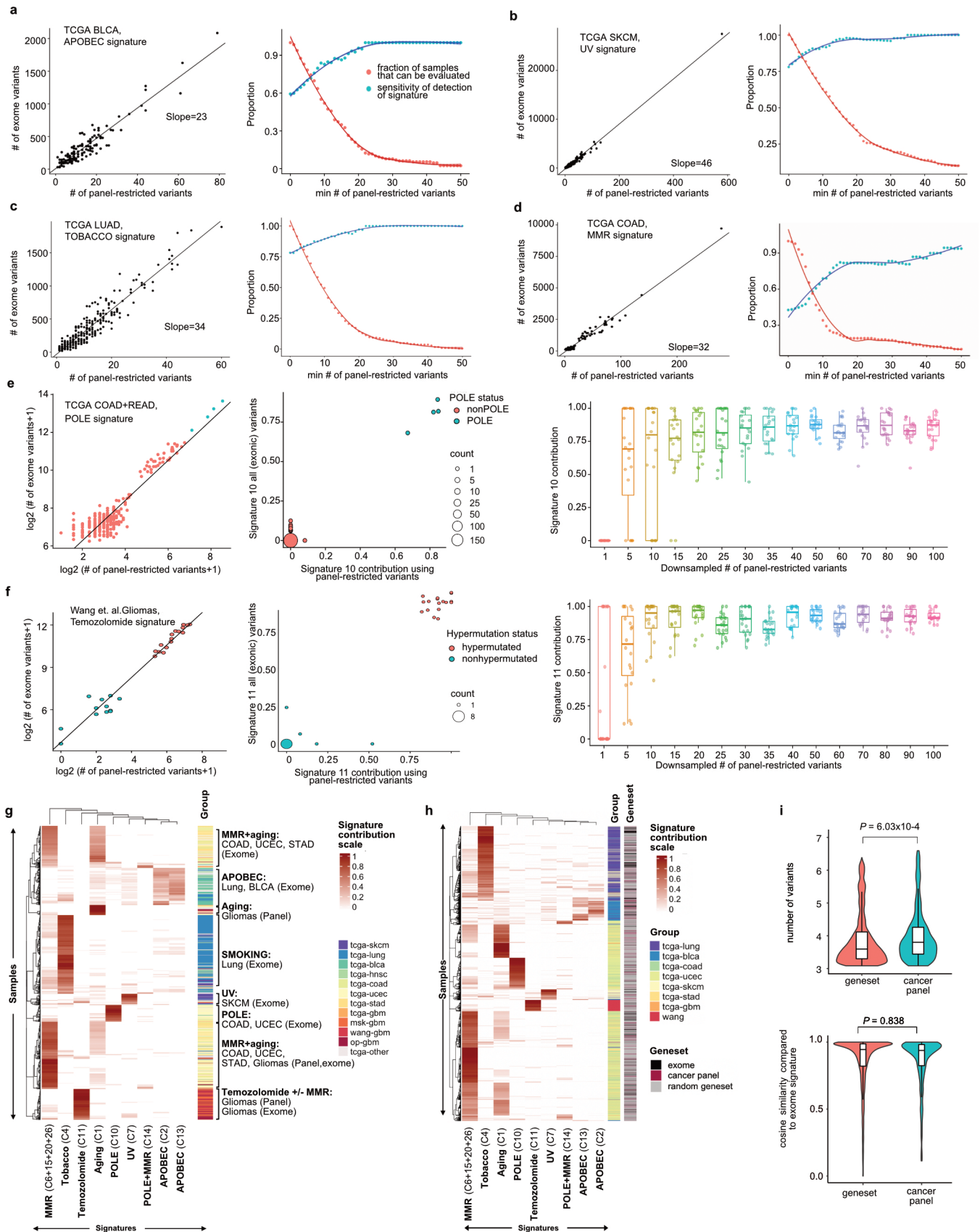
for hypermutation (13.8 mutations per Mb) was previously determined¹⁷. The frequency of hypermutation was similar in the three datasets (85 (5.2%) in DFCI-Profile; 29 (5.3%) in MSKCC-IMPACT; 444 (5.5%) in FMI). The median tumour mutation burden (TMB) in the combined datasets was 2.6 mutations per Mb (range, 0.0–781.3). Compared with non-hypermutated gliomas, hypermutated tumours showed atypical patterns of SNVs, consistent with abnormal mutational processes operating in these samples. Bars represent median and interquartile range for each dataset (right). HPI, homopolymer indels.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Integrated analysis of tumour mutation burden in hypermutated gliomas in the DFCI-Profile, MSKCC-IMPACT and FMI datasets. **a**, Distribution of TMB, homopolymer indels, MMR mutations, and SNV mutational spectrum according to molecular status of *IDH1/2*, 1p/19q co-deletion (1p/19q), gain of chromosome 7 and/or deletion of chromosome 10 (7gain/10del), and *MGMT* promoter methylation, histological grade, age at initial diagnosis, and history of prior treatment with alkylating agents or radiation therapy (the distinction between photon and proton therapy was not systematically captured) in the DFCI-Profile dataset ($n = 84$, data not shown for the single sample from other gliomas, *IDH1/2*-wt subgroup). **b**, Top, distribution of histomolecular groups in non-hypermuted and hypermutated gliomas from the combined sequencing dataset ($n = 2,173$). Bottom, distribution of molecular groups in de novo and post-treatment hypermutated gliomas from the DFCI-Profile dataset ($n = 85$) (annotation not available for the MSKCC-IMPACT set). **c**, Prevalence of hypermutation according to *MGMT* promoter methylation and *IDH1/2* mutation status in post-temozolomide gliomas from the DFCI-Profile dataset ($n = 150$). Two-sided Fisher's exact test. **d**, Number of temozolomide cycles according to *IDH1/2* mutation status in post-temozolomide diffuse gliomas from the DFCI-Profile dataset ($n = 211$ gliomas). Patients who received combined chemoradiation but no adjuvant temozolomide were included. Two-sided Wilcoxon rank-sum test. **e**, Boxplots of TMB in post-treatment hypermutated gliomas according to the number of

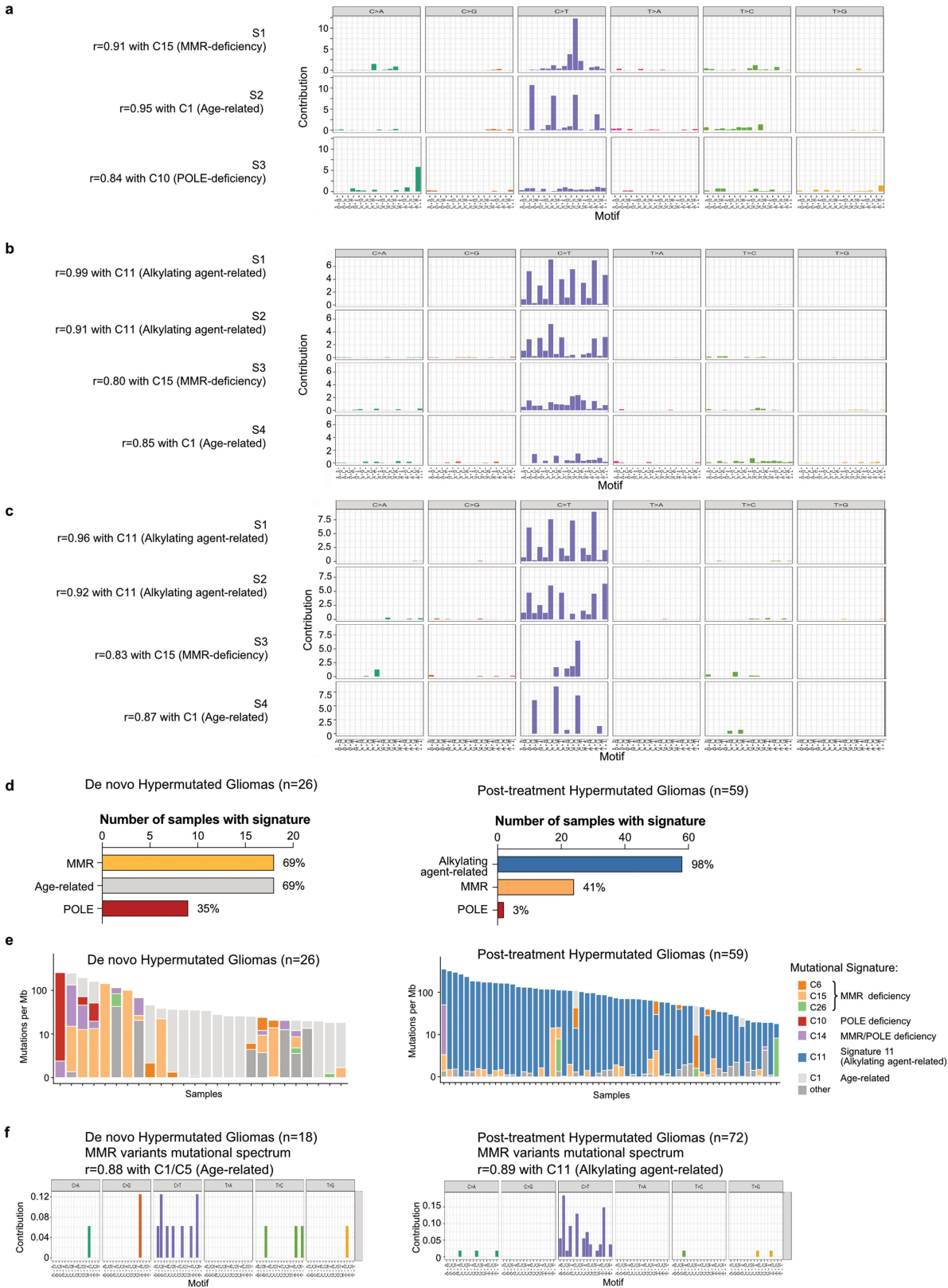
temozolomide cycles received before surgery. Kruskal–Wallis test and Dunn's multiple comparison test. **f**, TMB in recurrent gliomas according to treatments received before surgery. Patients who received multiple treatment modalities were excluded. Kruskal–Wallis test and Dunn's multiple comparison test. Boxes, quartiles; centre lines, median ratio for each group; whiskers, absolute range (**d–f**). **g**, Integrated analysis of the FMI dataset ($n = 8,121$ gliomas) depicting tumour mutation burden, the number of indels at homopolymer regions, and the SNV mutation spectrum detected in each tumour according to molecular status of *IDH1/2* and 1p/19q co-deletion (1p/19q), MSI status, and age at initial diagnosis. Dominant mutational signatures detected in hypermutated samples are depicted. The dotted line indicates the threshold for samples with a high mutation burden (8.7 mutations per Mb). **h**, Prevalence of hypermutation among molecularly defined subgroups in the FMI dataset ($n = 8,121$ gliomas). Chi-squared test. **i**, Dominant mutational signatures detected in hypermutated samples in the FMI dataset ($n = 8,121$ gliomas). Chi-squared test. **j**, Mutated genes and pathways enriched in hypermutated gliomas in the FMI dataset ($n = 8,121$). Enrichment was assessed using a permutation test to control for random effects of hypermutability in tumours with high TMB. **k, l**, Proportion of TMB^{high} versus TMB^{low} samples with mutations in selected DNA repair genes and glioma drivers (**e**) and in the MMR pathway (*MSH2*, *MSH6*, *MLH1* and *PMS2*; **f**). Permutation test; **** $P < 0.0001$, *** $P < 0.001$, ** $P < 0.01$; ns, not significant.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Validation of known hypermutation-associated signatures using TCGA datasets. Mutational signatures were predicted using exome-sequencing variants that overlapped with the panel-targeted regions, and then compared to previously published DeconstructSigs signature predictions based on all exome variants. The TCGA MC3 dataset was used to assess the detection of COSMIC mutational signatures associated with APOBEC (signatures 2 and 13), mismatch repair (signature 6), ultraviolet light (signature 7), POLE (signature 10), and tobacco (signature 4). Variant calls for 17 hypermutated and 12 non-hypermutated glioma exome-sequenced samples⁴ were used to assess temozolomide (signature 11) detection. **a**, Detection of APOBEC-associated mutational signature in TCGA BLCA samples ($n = 129$ out of 411 samples). **b**, Detection of ultraviolet-associated mutational signature in TCGA SKCM samples ($n = 237$ out of 466 samples). **c**, Detection of tobacco smoking-associated mutational signature in TCGA LUAD samples ($n = 250$ out of 513 samples). **d**, Detection of MMR-associated mutational signature in TCGA COAD ($n = 188$ out of 380 samples). **e**, Detection of POLE-associated mutational signature in TCGA COAD and READ samples ($n = 277$ out of 380 samples). **f**, Detection of temozolomide-associated mutational signature in ref.⁴ ($n = 29$). **g**, Unsupervised clustering of hypermutated samples. A total

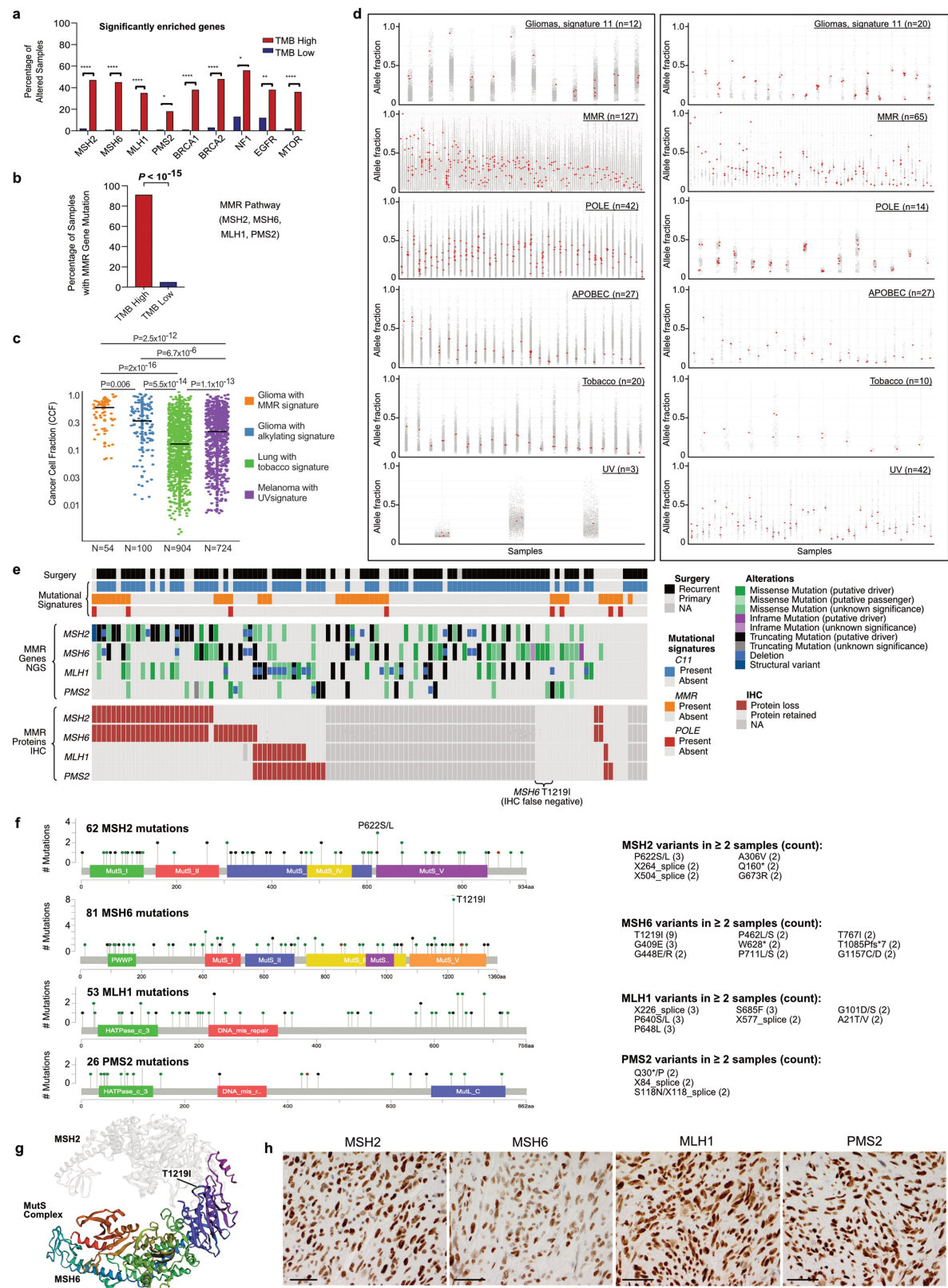
of 865 hypermutated tumour samples from exomes (pan-TCGA and Wang et al.⁴) and targeted panels (DFCI-Profile and MSK-IMPACT) were analysed for known hypermutation signatures (tobacco, UV, MMRD, POLE, TMZ, APOBEC). Samples and signatures underwent 2D hierarchical clustering based on Euclidean distance. **h**, Performance of cancer panel versus other genesets in mutational signature calling. We analysed 622 hypermutated tumour exomes (pan-TCGA and Wang et al.⁴, black) for their mutational signature contributions when restricted to variants from i) DFCI-Profile OncoPanel cancer panel genes (red), or ii) 9 randomly selected gene sets (grey) of similar total capture size to the cancer panel. For each sample, we assessed known hypermutation signatures for cancer panels and gene sets for which at least 20 single base substitutions were retained in the sample after restriction. Samples and signatures underwent 2D hierarchical clustering based on Euclidean distance. **i**, The violin plots represent the number of variants (top) and the cosine similarity of signature contributions (bottom) when using all exonic variants versus restriction to cancer panel or the 49 random gene sets. Boxes, quartiles; centre lines, median ratio for each group; whiskers, absolute range. Two-sided Welch's t -test.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Mutational signature analysis of primary and secondary hypermutated cohort (n = 111). **a**, Mutational signature analysis of newly diagnosed hypermutated gliomas in the DFCI-Profile dataset (n = 24). **b**, Mutational signature analysis of secondary hypermutated gliomas (samples in which hypermutation was detected in the recurrent tumour) in the DFCI-Profile dataset (n = 58). The novel COSMIC Signature 11-related signature (S2) was associated with 1p/19q co-deletion and lack of prior radiation therapy (66.7% of samples with high S2 versus 26.2% of samples with high S1 signature, Fisher $P = 0.016$). **c**, Mutational signature analysis of hypermutated gliomas from the MSKCC-IMPACT dataset (n = 29). **d**, Mutational signature analysis in de novo (hypermutated at first diagnosis, n = 26, left) and post-treatment hypermutated gliomas (hypermutation in a recurrent tumour, n = 59, right). Percentage of samples exhibiting the most common mutational signatures and their hypothesized causes are displayed. MMR, C6, C14, C15, C26;

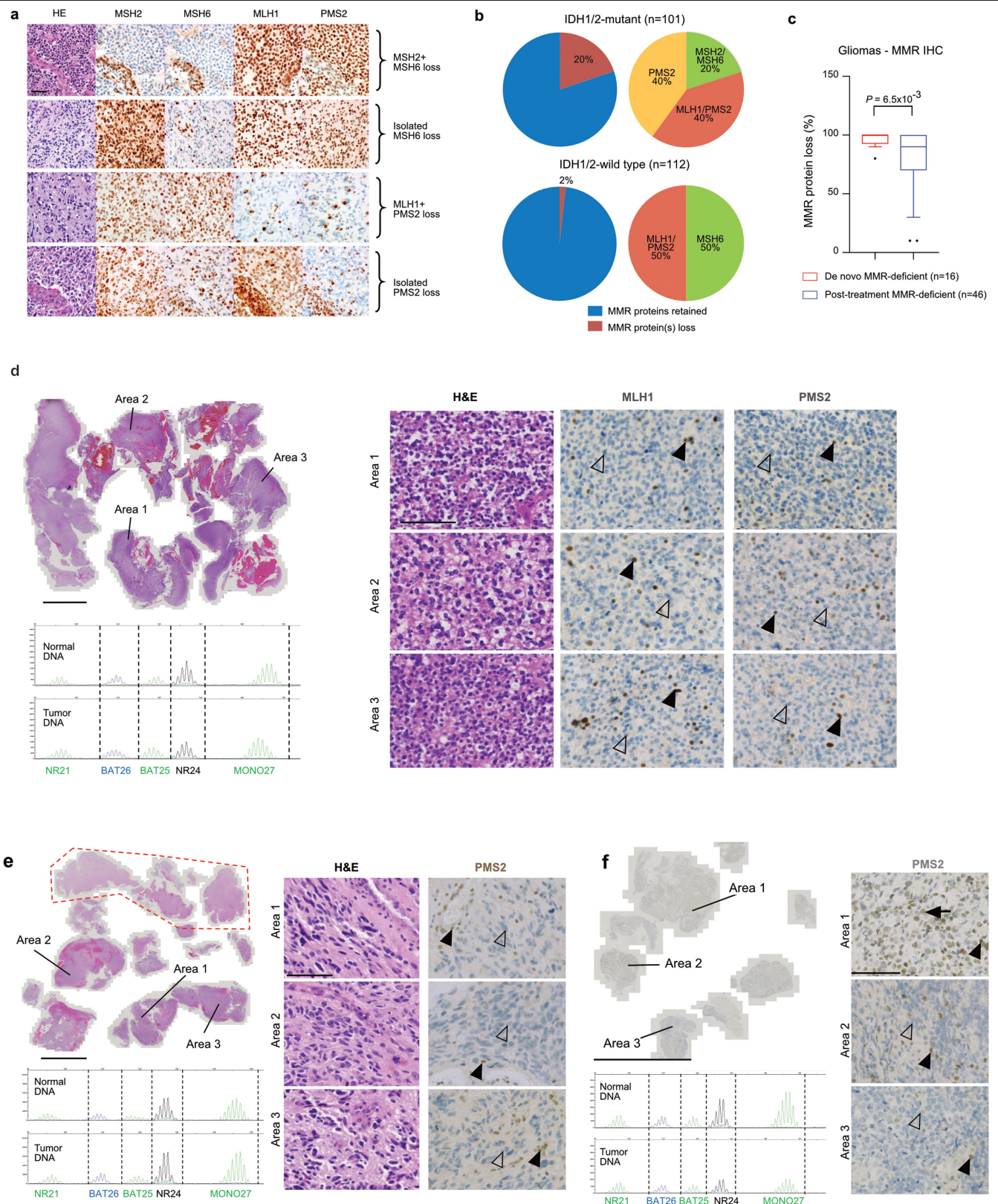
age-related, C1; POLE, C10, C14. Chi-squared test. **e**, Mutational signatures identified in individual de novo hypermutated gliomas (hypermutated at first diagnosis, n = 26, left) and post-treatment hypermutated gliomas (hypermutation in a recurrent tumour, n = 59, right). **f**, Mutational signature analysis of MMR variants in hypermutated gliomas from the DFCI-Profile and MSKCC-IMPACT datasets (n = 114). Ninety variants of the MMR genes *MSH2*, *MSH6*, *MLH1*, and *PMS2* were merged into two groups (de novo, n = 18; post-treatment, n = 72) according to the type of sample in which they were found and analysed for mutational signatures using a regression model (Rosenthal et al.⁵²). In each sample, only the MMR variant with the highest VAF was included, to limit the inclusion of possible passenger variants. For signature discovery in both cohorts (**a–c**), variants were analysed using the non-negative matrix factorization (NMF) method and correlated with known COSMIC mutational signatures¹⁴ using Pearson correlation.



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Characteristics of MMR molecular variants in hypermutated gliomas. **a, b**, Proportion of TMB^{high} versus TMB^{low} samples with mutations in selected DNA repair genes and glioma drivers (**a**) and in the MMR pathway (*MSH2*, *MSH6*, *MLH1* and *PMS2*) (**b**) in the merged DFCI-Profile/MSKCC-IMPACT dataset ($n = 2,173$). Permutation test; **** $P < 10^{-5}$, ** $P < 10^{-2}$, * $P < 0.05$. **c**, CCFs of MMR gene mutations in post-treatment hypermutated gliomas versus other hypermutated cancers in the FMI dataset. Horizontal line, median. Two-sided Wilcoxon rank-sum test with Benjamini–Hochberg correction. **d**, VAF distribution of mutations in post-treatment hypermutated gliomas, non-glioma MMR-deficient cancers (diverse histologies) and other non-glioma hypermutated samples (diverse histologies) from the TCGA and MSKCC-IMPACT datasets. Each dot represents a mutation found in an individual sample (represented vertically). MMR mutations are depicted in red. Left,

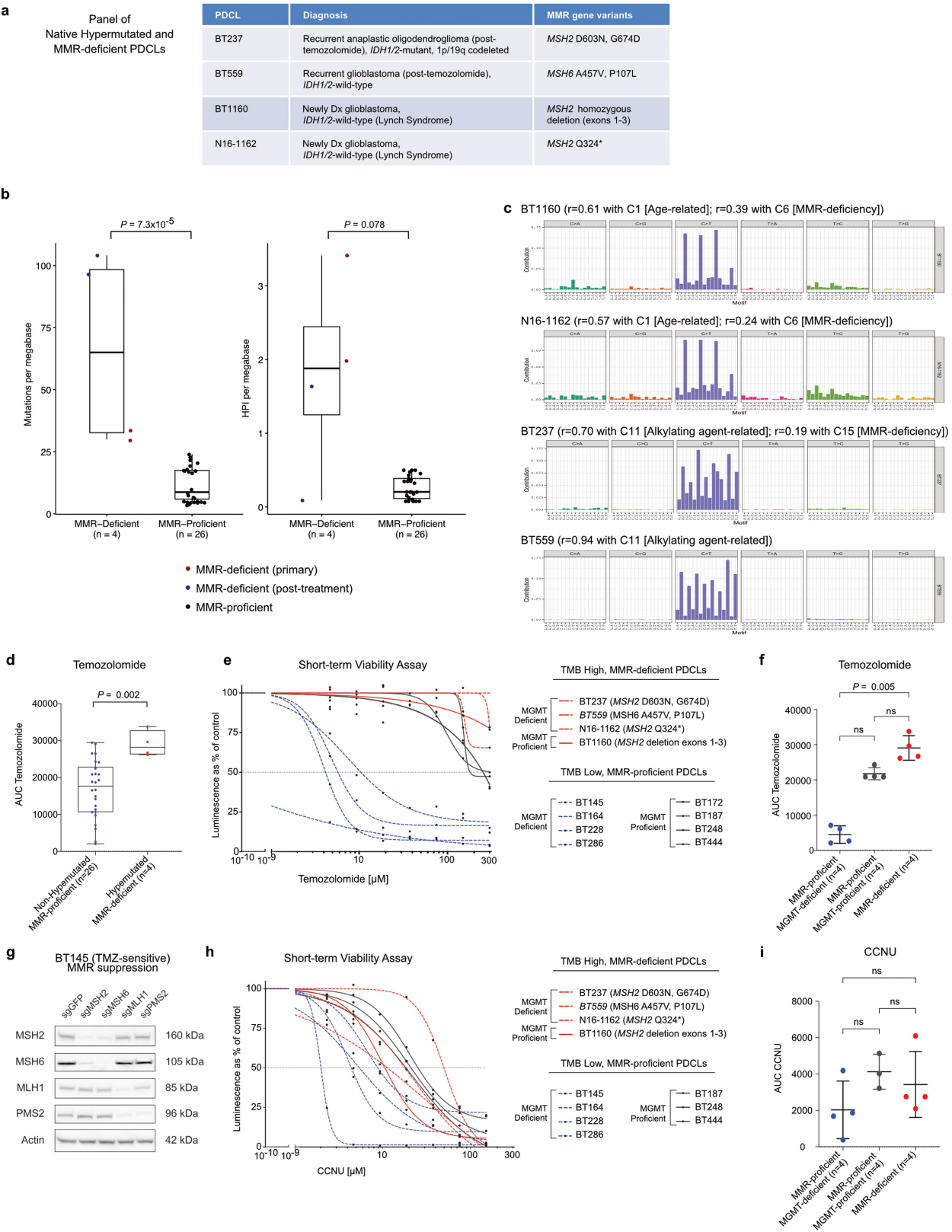
hypermutated samples from the pan-TCGA dataset; right, hypermutated samples from the MSKCC-IMPACT dataset. **e**, Integrated view of mutational signatures and MMR gene mutations and protein expression in hypermutated gliomas ($n = 114$). Tumours with the mutational hotspot MSH6(T1219I) (11.9% of post-treatment hypermutated gliomas) are highlighted. **f**, Mutation diagram of *MSH2*, *MSH6*, *MLH1*, and *PMS2* mutations found in hypermutated gliomas from the DFCI-Profile and MSKCC-IMPACT datasets ($n = 114$). The hotspot MSH6 missense variant p.T1219I was found in nine samples. **g**, Hotspot MSH6 p.T1219I variant mapped to the bacterial MutS 3D structure (PDB 5YK4). **h**, Representative immunohistochemistry (IHC) images of the MMR proteins MSH2, MSH6, MLH1 and PMS2 in a hypermutated glioblastoma with MSH6(T1219I) mutation. Three independent samples were stained. Scale bar, 100 μm .



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Results of MMR IHC screening in 213 consecutive recurrent gliomas and patterns of MMR protein expression loss in three de novo or post-treatment MMR-deficient gliomas. **a**, Recurrent patterns of MMR protein loss identified by IHC in gliomas. Scale bar, 50 μ m. **b**, Summary of MMR IHC screening results for 213 consecutive recurrent gliomas. All monocentric consecutive relapses of diffuse gliomas in adult patients following treatment with post-alkylating agents (surgery between 2009 and 2015) were included in the immunohistochemistry analysis. Further sequencing of samples in which MMR protein loss was identified showed hypermutation with MMR molecular defects in 18/19 (94.7%) samples. **c**, Percentage of tumour MMR protein loss in glioma samples with de novo ($n = 16$) or post-treatment ($n = 46$) MMR deficiency. Samples were scored by two pathologists in blinded fashion. Regional heterogeneity of MMR protein loss for the four MMR proteins MSH2, MSH6, MLH1, and PMS2 was scored as to the maximal percentage of protein loss among tumour cells for each sample (5% increments). Boxes, quartiles; centre lines, median ratio for each group; whiskers, absolute range, excluding outliers. Two-sided Wilcoxon rank-sum test. **d**, Clonal MMR deficiency in a de novo high-grade glioma. Top left, low magnification of haematoxylin and eosin (H&E) staining of the large surgical tumour pieces obtained from surgical resection. Right, high magnification in three tumour areas (H&E staining, MLH1 and PMS2 immunostaining) showing a highly cellular tumour with an oligodendroglial phenotype and a loss of expression of MLH1 and PMS2 in all tumour cells (open arrowheads). Normal cells have a maintained MLH1 and PMS2 expression (solid arrowheads). Bottom left, microsatellite testing via PCR amplification of five mononucleotide markers (BAT25, BAT26, NR21, NR24, and MONO27) showed the tumour to be MSS. Array CGH showed a homozygous deletion of the entire coding region of MLH1. Scale bars; top left, 5 mm; right, 100 μ m. **e**, Clonal MMR deficiency in a hypermutated post-treatment, *IDH1*-mutant glioblastoma. Top left,

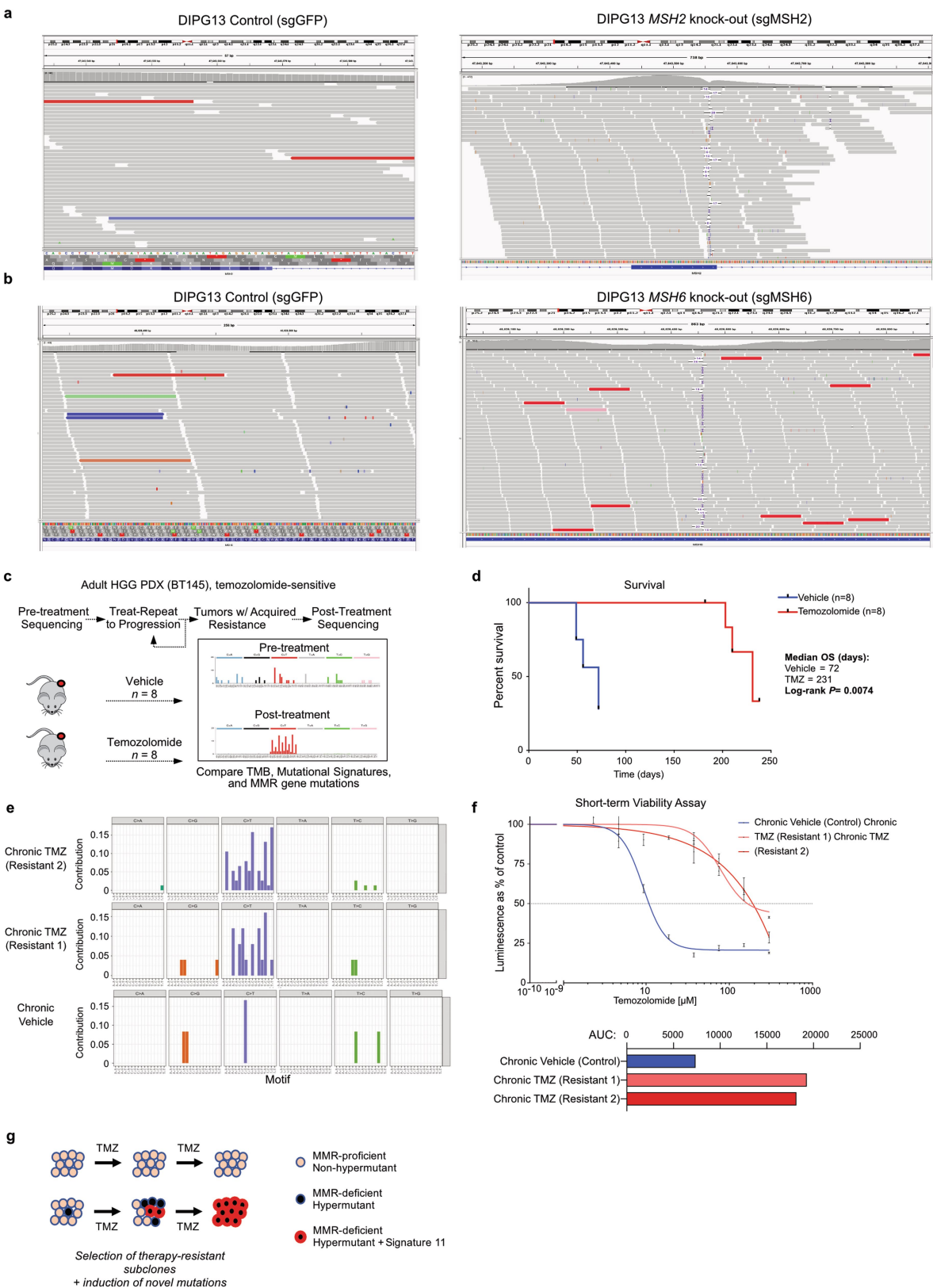
low-magnification image of H&E staining of tissue obtained from surgical resection, with three areas of tumour selected for images. Red dashed line delimits normal brain. Right, high-magnification images of H&E staining, showing highly cellular tumour and an astrocytic phenotype, and PMS2 IHC, showing loss of expression of PMS2 in all tumour cells (open arrowheads). Normal cells have maintained PMS2 expression (internal control, solid arrowheads). Bottom left, Microsatellite testing via PCR amplification of five mononucleotide markers (BAT25, BAT26, NR21, NR24, and MONO27) showed the tumour to be MSS. NGS showed a TMB of 120.1 per Mb and homopolymer indel burden of 3.8 per Mb, with contributions from temozolomide (90%) and MMR-deficiency (10%) mutational signatures. A missense (p.P648L) hotspot MLH1 mutation known to be pathogenic from patients with Lynch syndrome with a VAF of 0.73 and loss of heterozygosity was present in this case. Scale bars, top left, 5 mm; right 100 μ m. **f**, Subclonal MMR deficiency in a hypermutated post-treatment *IDH1*-mutant glioblastoma. Top left, low-magnification image of PMS2 immunostaining of the tumour pieces obtained from surgical resection. Right, high magnification images of three areas of PMS2 immunostaining showing heterogeneous PMS2 expression across the sample consistent with a subclonal tumour. Area 1 shows that PMS2 is retained in atypical tumour cells (arrow); area 2 is heterogeneous with loss (open arrowhead) in some but not all tumour cells; area 3 is an example of diffuse loss of expression in tumour cells (open arrowhead). Normal cells have a maintained PMS2 expression (solid arrowheads in all images). Bottom left, microsatellite analysis via PCR amplification of five mononucleotide markers (BAT25, BAT26, NR21, NR24, and MONO27) showed the tumour to be MSS. NGS showed a TMB of 236.5 per Mb and homopolymer indel burden of 2.3 per Mb, with 95% contribution of temozolomide mutational signature. Scale bars, top left 5 mm; right 100 μ m.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Characterization of high-grade glioma PDCLs and their sensitivity to temozolomide and CCNU. **a**, Clinico-molecular characteristics of four native newly diagnosed or recurrent glioma PDCL models harbouring hypermutation and MMR deficiency. **b**, Thirty glioma PDCLs, including four PDCLs derived from patients with de novo (BT1160, N16-1162, both established from patients with Lynch syndrome) or post-treatment (BT237, BT559) MMR deficiency were molecularly characterized using whole-exome sequencing. The panels show the tumour mutational burden (left) and homopolymer indel burden (right) in each model. Boxes, quartiles; centre lines, median ratio for each group; whiskers, absolute range. Two-sided Wilcoxon rank-sum test. **c**, Mutational signature analysis was performed in the PDCL models of constitutional and post-treatment MMR deficiency using the R package DeconstructSigs to estimate the contributions of mutational signatures using a regression model (Rosenthal et al.⁵²). For each PDCL, the contribution of the main COSMIC mutational signatures identified is

expressed as decimal. **d**, Boxplots of temozolomide AUC in non-hypermethylated versus hypermethylated PDCLs. Boxes, quartiles; centre lines, median ratio for each group; whiskers, absolute range. Two-sided Wilcoxon rank-sum test. **e, f**, A panel of 12 glioma PDCL models representing the different MGMT and MMR classes was selected and assessed for sensitivity to temozolomide in a short-term viability assay (**e**; dots represent means). The temozolomide AUC was compared between the three groups using a Kruskal–Wallis test and Dunn’s multiple comparison test (**f**; mean \pm s.d.). **g**, Western blot of the glioblastoma patient-derived cell line (BT145) in which the genes *MSH2*, *MSH6*, *MLH1* or *PMS2* have been knocked-out using the CRISPR–Cas9 system. **h, i**, A panel of 11 glioma PDCL models representing the different MGMT and MMR classes was selected and assessed for sensitivity to CCNU in a short-term viability assay (**h**; dots represent means). No CCNU data was available for the model BT172. The CCNU AUC was compared between the three groups using a Kruskal–Wallis test and Dunn’s multiple comparison test (**i**; mean \pm s.d.).

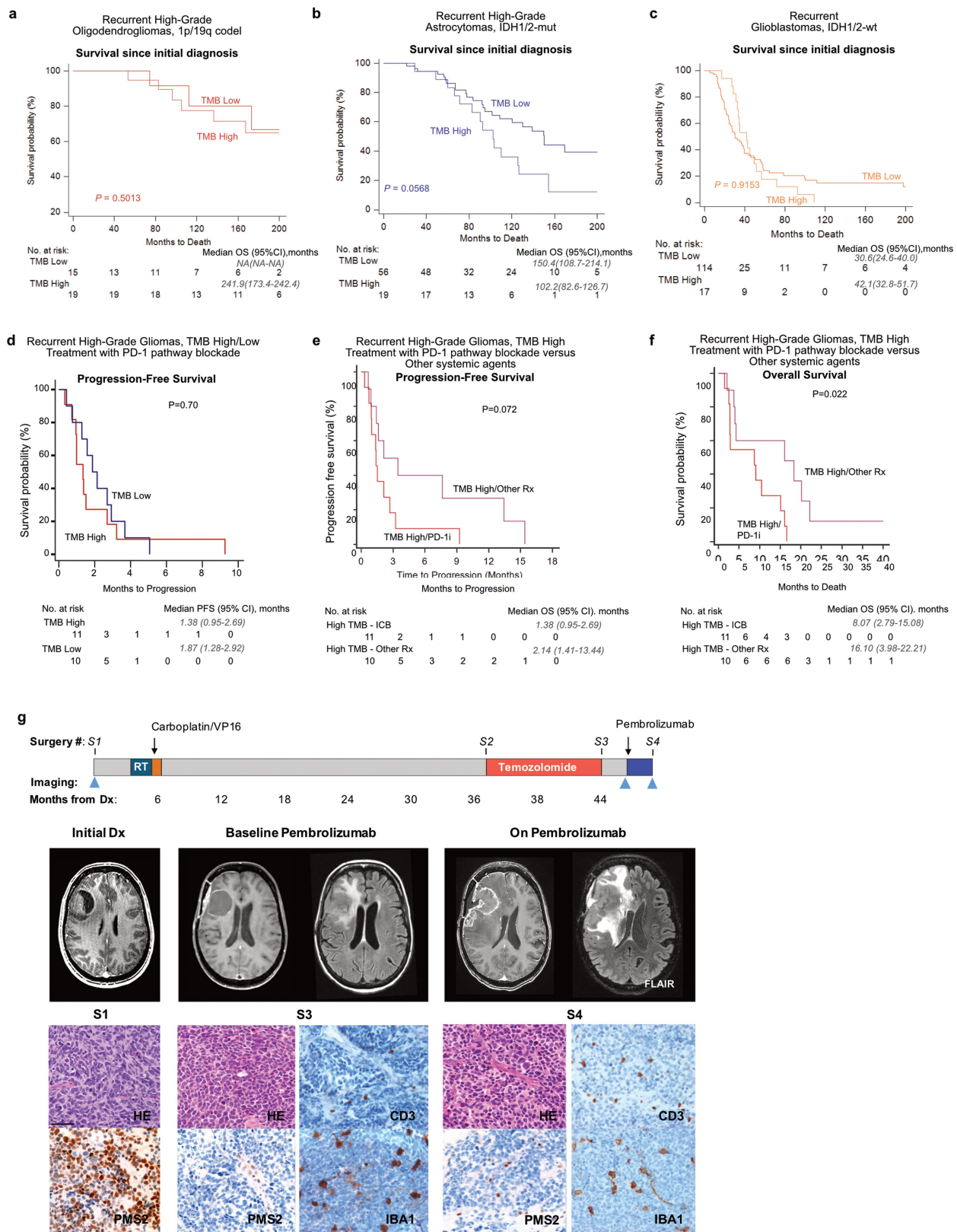


Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | MMR-deficient models of glioma, continued.

a, b, CRISPR–Cas9 *MSH2* and *MSH6* gene knockout in DIPG13 high-grade glioma cell line. **a**, Integrated genomics viewer (IGV) plots depicting *MSH2* reads in between the guide RNAs in the *MSH2* unedited line (sgGFP, left) and the *MSH2* CRISPR knockout line (right) confirming knockout in the *MSH2* edited line. **b**, IGV plots depicting *MSH6* reads in between the guide RNAs in the *MSH6* unedited line (sgGFP, left) and the *MSH6* CRISPR knockout line (right) confirming knockout in the *MSH6* edited line. **c**, Overview of in vivo temozolomide resistance study. Treatment of subcutaneous BT145 PDX-bearing animals was initiated at a volume of 100 mm³ and eight nude mice per group were randomized to 12 mg/kg/day temozolomide or vehicle for five consecutive days per 28-day cycle. Mice were treated until tumours reached a volume of 1,500 mm³, and tumours were sequenced to identify mutations and mutational signature. **d**, Survival of mice with BT145 xenografts (*n* = 8 mice per group) during treatment with vehicle (blue) or temozolomide (red). Two-sided log-rank test. **e**, Unique variants found in three sequenced BT145 tumours (two temozolomide-treated, and one vehicle-treated) were analysed for

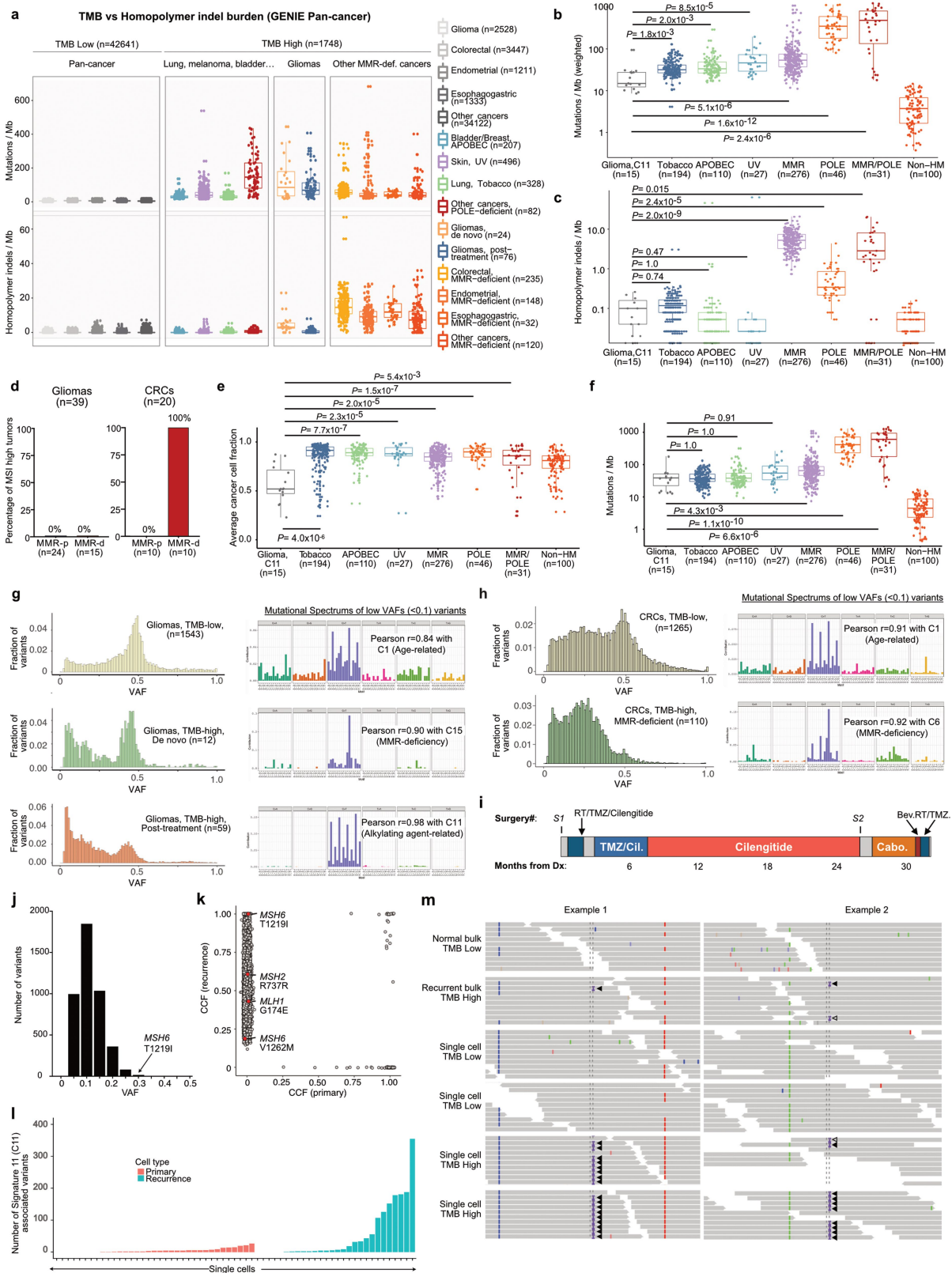
correlation with known mutational signatures. COSMIC Signature 11 was found in the two temozolomide-treated tumours. Mutational signatures could not be called in the vehicle-treated tumour (too few variants). After filtering of truncal variants common to all tumours, the two temozolomide-treated tumours shared only four variants, including an *MSH6*(T1219I) mutation and three noncoding variants. **f**, BT145 xenografts chronically treated with vehicle (*n* = 1) or temozolomide (*n* = 2) were removed, dissociated and cultured in serum-free medium to establish cell lines. After three passages in culture, sensitivity to temozolomide was assessed. The results of the short-term viability assays (mean ± s.e.m.) and temozolomide AUC of each cell line are depicted. **g**, Model of acquired hypermutation with mutational signature 11 in gliomas. Top, MMR-proficient cells repair TMZ damage and do not develop signature 11. Resistance in these cells is mediated by non-MMR pathways (for example, MGMT expression). Bottom, TMZ induces and/or selects resistant subclonal MMR-deficient cells. Further TMZ exposure produces accumulation of mutations at specific trinucleotide contexts, detected as hypermutation with signature 11.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Extended outcome data. a–c, Survival of patients with recurrent high-grade glioma (WHO grade III or IV) from the time of initial diagnosis according to TMB status (solid curves, TMB^{low}; dotted curves, TMB^{high}). The curves include 240 recurrent samples from DFCI-Profile with available survival data from initial diagnosis. Two-sided log-rank test. **a,** Survival of patients with recurrent high-grade 1p/19q co-deleted oligodendroglioma from the time of initial diagnosis. **b,** Survival of patients with recurrent high-grade *IDH1/2*-mutant astrocytoma from the time of initial diagnosis. **c,** Survival of patients with recurrent *IDH1/2* wild-type glioblastoma from the time of initial diagnosis. **d,** PFS of 11 patients with hypermutated and MMR-deficient glioma who were treated with PD-1 blockade (single-agent or in combination with bevacizumab, red curve). A cohort of patients with non-hypermutated glioma who were treated with PD-1 blockade is depicted as control (*n* = 10, best matches according to diagnosis, primary versus recurrent status, and prior treatments, blue curve). A two-sided log-rank test is used.

e, f, PFS (**e**) and OS (**f**) of 11 patients with hypermutated and MMR-deficient glioma who were treated with PD-1 blockade (red curves). A cohort of hypermutated patients treated with other systemic agents is depicted as control (best matches according to diagnosis, primary vs recurrent status, and prior treatments were selected from the cohort of sequenced gliomas, purple curves). Two-sided log-rank test. Clinical and histomolecular characteristics of patients from both cohorts are provided in Supplementary Table 7. **g,** Lack of immune response following PD1 blockade (pembrolizumab) in a patient with post-treatment hypermutated MMR-deficient glioblastoma. Top, timeline; middle, MRI images; bottom, H&E images and IHC for PMS2 expression and tumour infiltration with CD3-positive T cells and IBA1-positive macrophages in the primary (S1), recurrent pre-pembrolizumab (S3) and recurrent post-pembrolizumab (S4) tumours. The tumour acquired a focal *PMS2* two-copy deletion, protein loss, and hypermutation in the post-temozolomide recurrent tumour (S3). Scale bar, 50 µm.



Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Molecular characteristics of hypermutated gliomas.

a, Pan-cancer analysis of TMB and homopolymer indel burden in the GENIE dataset ($n = 44,389$). Tumour samples from the GENIE dataset (v6.1) were analysed for mutational and homopolymer indel burden. Statistical comparisons between groups are provided in Supplementary Table 6. **b**, TMB in hypermutated gliomas (post-treatment) versus MMR-deficient cancers and other hypermutated cancers from the TCGA and Wang et al.⁴ exome datasets ($n = 798$). Two-sided Wilcoxon rank-sum test with Bonferroni correction. **c**, Pan-cancer analysis of the homopolymer indel burden in hypermutated gliomas (post-treatment) versus MMR-deficient cancers and other hypermutated cancers from the TCGA and Wang et al.⁴ exome datasets ($n = 798$). **d**, Results of MSI analysis using the standard pentaplex assay in glioma ($n = 39$) and CRC samples ($n = 19$) according to MMR status (MMR-d, MMR deficient; MMR-p, MMR-proficient). **e**, Pan-cancer analysis of cancer cell fractions in hypermutated gliomas (post-treatment) versus MMR-deficient cancers and other hypermutated cancers from the TCGA and Wang et al.⁴ exome datasets ($n = 798$). Two-sided Wilcoxon rank-sum test with Bonferroni correction. **f**, Weighted TMB in hypermutated gliomas (post-treatment) versus MMR-deficient cancers and other hypermutated cancers from the TCGA and Wang et al.⁴ exome datasets ($n = 798$). The weighted TMB was calculated by weighing each individual mutation to its cancer cell fraction. Two-sided Wilcoxon rank-sum test with Bonferroni correction. **g**, Distribution of VAFs (left) and mutation spectrum analysis of low-allelic frequency variants (<0.1 , right) in TMB^{low} gliomas ($n = 1,543$, top), de novo hypermutated gliomas with MMR deficiency mutational signature ($n = 12$, middle), and post-treatment hypermutated gliomas ($n = 59$, bottom) from the DFCI-Profile dataset. **h**, Distribution of VAFs (left) and mutation signature analysis of low-allelic frequency variants (<0.1 , right) in TMB^{low} CRCs ($n = 1,265$, top) and TMB^{high} CRCs with MMR deficiency mutational signature ($n = 110$, bottom) from the GENIE

dataset. **i**, Clinical timeline for the patient with hypermutated glioblastoma with an MSH6(T1219I) mutation in whom bulk and single-cell WGS was performed. **j**, Distribution of VAFs of mutations in the recurrent bulk sample. The median VAF in the recurrent sample was 0.11. The MSH6(T1219I) mutation had the 18th-highest VAF out of 4,350 coding mutations. **k**, Cancer cell fractions (CCFs) of mutations in the primary and recurrent tumour bulk samples. Each dot represents a coding mutation. The horizontal and vertical axes are estimated clonal frequency for each mutation in the primary and recurrent samples, respectively. Mutations of the four main MMR genes are depicted in red. **l**, Mutational spectra in 35 cells from the primary tumour (orange) and 28 from the recurrent tumour (green) submitted to scWGS sequencing (1×). Mutational signature analysis showed a strong contribution of mutational signature 11 in hypermutated cells from the recurrent tumour. **m**, Representative IGV plots ($n = 2$ distinct genomic segment for each sample) of microsatellite insertions in the normal (TMB low) and recurrent (TMB high) bulk samples and recurrent TMB low ($n = 2$) and TMB high ($n = 2$) single cells. Solid arrowheads represent microsatellite insertions phased with a flanking heterozygous SNP allele. Open arrowheads represent microsatellite insertions for which the reads do not reach the flanking heterozygous SNP allele. Both hypermutated single cells showed multiple phased microsatellite insertions consistent with a true somatic microsatellite mutation. In general, a few reads with similar microsatellite insertions correctly phased with the same flanking heterozygous SNP allele were found in the recurrent bulk, but not in the normal bulk or non-hypermutated cells. For **a–c**, **e**, **f**, biological subgroups were identified on the basis of mutational burden, dominant signature and histology. For **b**, **c**, **e**, **f**, 100 non-hypermutated samples were randomly selected as controls. For all box plots: boxes, quartiles; centre lines, median ratio for each group; whiskers, absolute range, excluding outliers. RT, radiation therapy; Cil, cilengitide; Cabo, cabozantinib; Bev, bevacizumab.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☐ ☒ A description of all covariates tested
- ☐ ☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☐ ☒ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No software used for data collection

Data analysis

For WES analyses, we used the CGA WES Characterization pipeline developed at the Broad Institute to call, filter and annotate somatic mutations and copy number variation. The pipeline employs the following tools: MuTect[1], ContEst[2], Strelka[3], Orientation Bias Filter[4], DeTiN [5], AllelicCapSeg[6], MAFFoNFilter[7], RealignmentFilter, ABSOLUTE[8], GATK[9], PicardTools[10], Variant Effect Predictor [11], Oncotator [12].

1. MuTect1: Cibulskis, K, Lawrence, MS & Carter, SL. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature* ... (2013). doi:10.1038/nbt.2514
2. ContEst: Cibulskis, K. et al. ContEst: estimating cross-contamination of human samples in next-generation sequencing data. *Bioinformatics* 27, 2601–2 (2011).
3. Strelka: Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 28, 1811–7 (2012).
4. Orientation Bias Filter (oxoG, FFPE): Costello, M. et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* 41, e67 (2013).
5. DeTiN: Taylor-Weiner, A. et al. DeTiN: overcoming tumor-in-normal contamination. *Nat Methods* 15, 531–534 (2018).
6. AllelicCapSeg: Landau, D. A. et al. Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* 152, 714–26 (2013).
7. MAFFoNFilter: Lawrence, M. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495 (2014).
8. ABSOLUTE: Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* 30, 413–21 (2012). doi: 10.1038/nbt.2203.
9. GATK (Mutect2, somatic CNV): McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–303 (2010).
10. Picard Tools: https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.1.0/picard_fingerprint_CrosscheckFingerprints.php
https://software.broadinstitute.org/gatk/documentation/tooldocs/4.0.0.0/picard_analysis_CollectMultipleMetrics.php

11. Variant Effect Predictor: McLaren, W. et al. The Ensembl Variant Effect Predictor. *Genome Biol.* 17, 122 (2016).
12. Oncotator: Ramos, A. H. et al. Oncotator: cancer variant annotation tool. *Hum. Mutat.* 36, E2423–9 (2015).

Additional tools used for PDX analyses, mutational signature analyses and statistical analyses included:

1. SAMtools (1.7): Li, H. et al A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 2011;27:2987-93
2. BWA-MEM (0.7.17): <https://github.com/lh3/bwa>
3. Disambiguate (ngs_disambiguate-1.0): <https://github.com/AstraZeneca-NGS/disambiguate>
4. Genemapper Software (5): <https://www.thermofisher.com/order/catalog/product/4475073#/4475073>
5. SomaticSignatures (3.1): <https://bioconductor.org/packages/release/bioc/html/SomaticSignatures.html>
6. DeconstructSigs (1.8): <https://github.com/raerose01/deconstructSigs>
7. STATA (v14.2): <https://www.stata.com>
8. MedCalc (19.0.3): <https://www.medcalc.org>
9. Graphpad Prism (8): <https://www.graphpad.com/scientific-software/prism/>

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Clinical and sequencing data from 1495 samples from the DFCI-Profile and 545 samples from the MSKCC-IMPACT datasets are publicly available (GENIE v4.1: <https://www.synapse.org/>). All data for samples from the GENIE v6.1 and TCGA pan-cancer datasets are publicly available. Data for samples from the FMI dataset are not publicly available.

Detailed clinical annotation for the DFCI-Profile and MSKCC-IMPACT cohorts is provided in supplementary table 1.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical test was used to determine the sample size. We systematically collected data from 10,294 glioma samples from three large independent datasets. Based on prior literature (prevalence of hypermutation of 2-5% in gliomas), we hypothesized that this sample size would provide enough power (200-500 hypermutated samples) for clinical and molecular association studies.
Data exclusions	47 samples for which the clinical diagnosis of glioma could not be confirmed (other histology or possible non-tumor sample) and 5 samples with missing clinical annotation were excluded from all analyses. Exclusion criteria were pre-established.
Replication	In vivo experiments were performed using groups of 8 mice per group. All mice were included in the analysis. In vitro sensitivity assays were replicated in at least 3 independent experiments. All experiments that were technically valid were included in the analysis.
Randomization	For in vivo experiments, mice were randomized.
Blinding	No blinding was performed. Blinding was not relevant to our study.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

For immunohistochemistry, the following antibodies were used: mouse monoclonal anti-ATRX (Bio SB, clone BSB-108, BSB3296, 1:100), mouse monoclonal anti-IDH1 R132H (Dianova, clone H09, DIA-H09, 1:100), rabbit monoclonal anti-CD3 (Roche, clone 2GV6, 790-4341, prediluted), rabbit polyclonal anti-IBA1 (Wako, W1W019-19741, 1:500), mouse monoclonal anti-MLH1 (Roche, clone M1, 790-4535, prediluted), mouse monoclonal anti-MSH2 (Roche, clone G219-1129, 760-4265, prediluted), mouse monoclonal anti-MSH6 (Roche, clone 44, 760-4455, prediluted), rabbit monoclonal anti-PMS2 (Roche, clone EPR3947, 760-4531, prediluted). For immunohistochemistry performed at BWH, the following antibodies were used: mouse monoclonal anti-MLH1 (Leica, clone ES05, MLH1-L-CE, 1:75), mouse monoclonal anti-MSH2 (Merck Millipore, clone Ab-2-FE11, NA27, 1:200), mouse monoclonal anti-MSH6 (Leica, clone PU29, MSH6-L-CE, 1:50), mouse monoclonal anti-PMS2 (Cell Marque, MRQ-28, 288M-14-ASR, 1:100). For immunoblotting, the following antibodies were used: mouse monoclonal anti-MGMT (Millipore, MT3.1, MAB16200, 1:500), mouse monoclonal anti-MSH2 (Calbiochem, FE11, NA27, 1:1000), mouse monoclonal anti-MSH6 (Biosciences, 44, 610918, 1:500), mouse monoclonal anti-MLH1 (Cell Signaling, 4C9C7, 3515, 1:500), mouse monoclonal anti-PMS2 (BD Biosciences, A16-4, 556415, 1:1000), mouse monoclonal anti-beta-actin (Sigma, AC-74, A2228, 1:10000).

Validation

All antibodies used are commercially available and were validated by the manufacturer.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

All patient-derived cell lines (PDCls) and xenografts (PDXs) with a name starting with "BT" were established from tumors resected at Brigham and Women's Hospital and Boston Children's Hospital (Boston, MA). SU-DIPG-XIII (DIPG13) cells were provided by Dr. Michelle Monje at Stanford University.

Authentication

The identity of all cell lines established were confirmed by short tandem repeat assay.

Mycoplasma contamination

All cell lines were tested for the absence of mycoplasma.

Commonly misidentified lines
(See [ICLAC](#) register)

No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

Eight-week-old NU/NU male mice (Charles River).

Wild animals

The study did not involve wild animals.

Field-collected samples

The study did not involve samples collected from the field.

Ethics oversight

All in vivo studies were performed in accordance with Dana-Farber Cancer Institute animal facility regulations and policies under the protocol #09-016.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

We systematically analyzed clinical data and somatic tumor variants identified through targeted next-generation sequencing (NGS) panels of 1,628 gliomas sequenced between June 2013 and November 2018 as part of a large institutional prospective profiling program (DFCI-Profile). All samples were assigned to a molecular subgroup based on histopathology, mutational status of IDH1 and IDH2 genes, and whole-arm co-deletion of chromosomes 1p and 19q (1p/19q co-deletion). A total of 545 independent samples from the GENIE project (a repository of genomic data obtained during routine clinical care at international institutions) were also identified and assigned to molecular subgroups. The combined sequencing set comprised 2,173 glioma samples, which included a broad spectrum of newly-diagnosed and recurrent glioma types including IDH1/2 wild-type

glioblastomas (1234 tumors, 56.8%), IDH1/2-mutant gliomas (640, 29.5%), and other rare subtypes of IDH1/2 wild-type gliomas (299, 13.8%). In addition, 247 gliomas collected at one site between 2009 and 2017 were analyzed for targeted protein expression using immunohistochemistry.

Recruitment

Genomic testing was ordered by the pathologist or treating physician as part of routine clinical care to identify relevant genomic alterations that could potentially inform diagnosis and treatment decisions. All patients undergoing genomic testing signed a clinical consent form, permitting to return results from clinical sequencing. No systematic bias likely to impact results were known at the time of data analysis.

Ethics oversight

The study, including the consent procedure, was approved by the institutional ethics committees (10-417/11-104/17-000; WIRB, Puyallup WA).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Paracrine orchestration of intestinal tumorigenesis by a mesenchymal niche

<https://doi.org/10.1038/s41586-020-2166-3>

Received: 29 November 2018

Accepted: 3 February 2020

Published online: 1 April 2020

 Check for updates

Manolis Roulis^{1,12}✉, Aimilios Kaklamanos^{2,12}, Marina Scherthanner¹, Piotr Bielecki¹, Jun Zhao^{1,3,4}, Eleanna Kaffe¹, Laura-Sophie Frommelt¹, Rihao Qu^{1,3,4}, Marlene S. Knapp¹, Ana Henriques², Niki Chalkidi², Vasiliki Koliarakis², Jing Jiao⁵, J. Richard Brewer¹, Maren Bacher¹, Holly N. Blackburn¹, Xiaoyun Zhao⁶, Richard M. Breyer^{7,8}, Vassilis Aidinis², Dhanpat Jain⁴, Bing Su⁶, Harvey R. Herschman⁵, Yuval Kluger^{3,4,9}, George Kollias^{2,10,13}✉ & Richard A. Flavell^{1,11,13}✉

The initiation of an intestinal tumour is a probabilistic process that depends on the competition between mutant and normal epithelial stem cells in crypts¹. Intestinal stem cells are closely associated with a diverse but poorly characterized network of mesenchymal cell types^{2,3}. However, whether the physiological mesenchymal microenvironment of mutant stem cells affects tumour initiation remains unknown. Here we provide *in vivo* evidence that the mesenchymal niche controls tumour initiation in *trans*. By characterizing the heterogeneity of the intestinal mesenchyme using single-cell RNA-sequencing analysis, we identified a population of rare pericryptal *Ptgs2*-expressing fibroblasts that constitutively process arachidonic acid into highly labile prostaglandin E₂ (PGE₂). Specific ablation of *Ptgs2* in fibroblasts was sufficient to prevent tumour initiation in two different models of sporadic, autochthonous tumorigenesis. Mechanistically, single-cell RNA-sequencing analyses of a mesenchymal niche model showed that fibroblast-derived PGE₂ drives the expansion of a population of Sca-1⁺ reserve-like stem cells. These express a strong regenerative/tumorigenic program, driven by the Hippo pathway effector Yap. *In vivo*, Yap is indispensable for Sca-1⁺ cell expansion and early tumour initiation and displays a nuclear localization in both mouse and human adenomas. Using organoid experiments, we identified a molecular mechanism whereby PGE₂ promotes Yap dephosphorylation, nuclear translocation and transcriptional activity by signalling through the receptor Ptger4. Epithelial-specific ablation of *Ptger4* misdirected the regenerative reprogramming of stem cells and prevented Sca-1⁺ cell expansion and sporadic tumour initiation in mutant mice, thereby demonstrating the robust paracrine control of tumour-initiating stem cells by PGE₂–Ptger4. Analyses of patient-derived organoids established that PGE₂–PTGER4 also regulates stem-cell function in humans. Our study demonstrates that initiation of colorectal cancer is orchestrated by the mesenchymal niche and reveals a mechanism by which rare pericryptal *Ptgs2*-expressing fibroblasts exert paracrine control over tumour-initiating stem cells via the druggable PGE₂–Ptger4–Yap signalling axis.

Mesenchymal cells are localized in tight association with stem cells in crypts, separated from them by a layer of extracellular matrix less than a 1 µm in thickness (Extended Data Fig. 1a, b). To investigate the heterogeneity of intestinal mesenchymal cells and identify specific pathways that could control stem-cell dynamics, we performed

single-cell RNA-sequencing (RNA-seq) analysis in the mouse intestinal mesenchyme. By sequencing 3,179 non-epithelial, non-immune cells we identified all major mesenchymal cell types of the lamina propria, including vascular and lymphatic endothelial cells, pericytes, smooth muscle cells and glial cells (Fig. 1a). Our analyses also revealed

¹Department of Immunobiology, Yale University School of Medicine, New Haven, CT, USA. ²Biomedical Sciences Research Center 'Alexander Fleming', Vari, Greece. ³Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. ⁴Department of Pathology, Yale University School of Medicine, New Haven, CT, USA. ⁵Department of Molecular and Medical Pharmacology, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA. ⁶Shanghai Institute of Immunology, Department of Microbiology and Immunology, Shanghai Jiao Tong University School of Medicine, Shanghai, People's Republic of China. ⁷Department of Veterans Affairs, Tennessee Valley Health Authority, Nashville, TN, USA. ⁸Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. ⁹Applied Mathematics Program, Yale University, New Haven, CT, USA. ¹⁰Department of Physiology, Medical School, National and Kapodistrian University of Athens, Athens, Greece. ¹¹Howard Hughes Medical Institute, Yale University School of Medicine, New Haven, CT, USA. ¹²These authors contributed equally: Manolis Roulis, Aimilios Kaklamanos. ¹³These authors jointly supervised this work: George Kollias, Richard A. Flavell. ✉e-mail: emmanouil.roulis@yale.edu; kollias@fleming.gr; richard.flavell@yale.edu

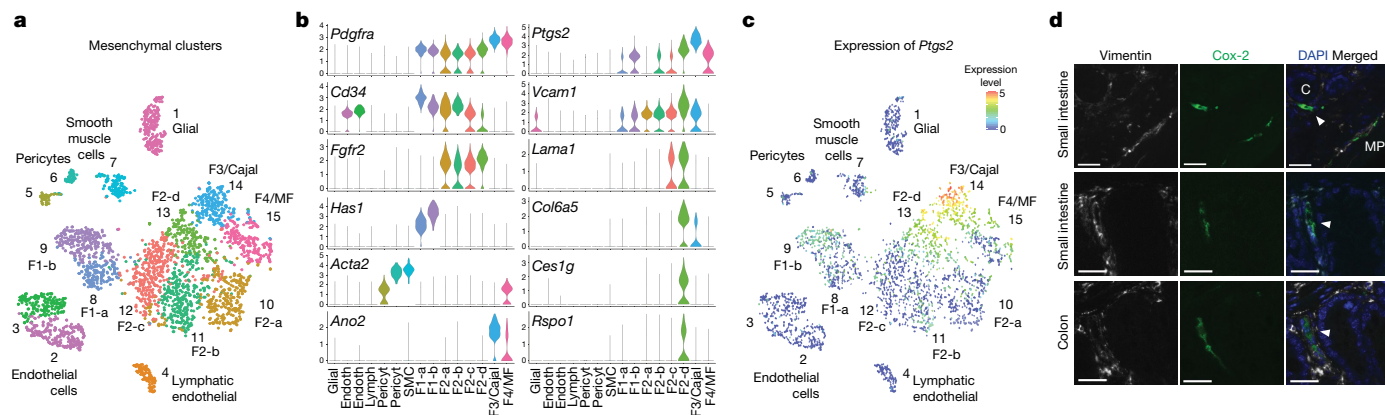


Fig. 1 | Single-cell analyses of the intestinal mesenchyme reveal a rare fibroblast population that expresses *Ptgs2* and its protein product Cox-2, located under the crypts. a–c, Single-cell RNA-seq of 3,179 mesenchymal cells from the normal mouse colon. **a**, *t*-distributed stochastic neighbour embedding (*t*-SNE) plot with clustering results. F, fibroblasts. MF, myofibroblasts. **b**, Violin plots showing the entire range of mesenchymal marker gene expression levels per single cell in each cluster. Endoth.,

endothelial; lymph, lymphatic; pericyt, pericytes; SMC, smooth muscle cells. **c**, *Ptgs2* expression levels per single cell visualized by *t*-SNE plot. **d**, Immunostaining for Cox-2 (green) and vimentin (white) in the normal mouse ileum and colon. Cox-2-expressing fibroblasts are located under the crypt (C) epithelium (white arrowheads) and in the muscularis propria (MP). Results are representative of three independent experiments. Scale bars, 20 μ m.

the existence of four different fibroblast populations—which we designate F1 to F4—all characterized by expression of *Pdgfra* (Fig. 1b); F1 and F2 cells are *Pdgfra*^{low}, whereas F3 and F4 cells are *Pdgfra*^{high} (Fig. 1b). A similar division of fibroblasts into *PDGFRA* high and low populations was found in a single-cell dataset³ of the human colonic mesenchyme (Extended Data Fig. 1h). Confocal and two-photon imaging in *Pdgfra*^{eGFP/+}-knockin mice⁴ confirmed the presence of *Pdgfra*^{high} and *Pdgfra*^{low} fibroblasts and revealed a distinct localization of these two populations in the adult intestine (Extended Data Fig. 2a, b). A dichotomous and compartmentalized expression of *Pdgfra* was observed in the intestinal mesenchyme throughout postnatal development and on embryonic day 15, when *Pdgfra*^{high} cells are associated with early villus formation, whereas *Pdgfra*^{low} cells occupy the rest of the mesenchyme (Extended Data Fig. 2c). Among *Pdgfra*^{low} cells, F1 fibroblasts comprise two subsets (F1a and F1b), both marked by expression of *Cd34* and *Has1* (Fig. 1b). Population F2 expresses *Cd34* and *Fgfr2* and comprises four subsets (F2a–F2d) occupying diverse niches in the intestine, as shown in *Fgfr2*^{mCherry}-knockin mice⁵ (Fig. 1b, Extended Data Fig. 2d). Among *Pdgfra*^{high} cells, F3 cells express the Cajal-cell marker *Ano2* whereas F4 cells express *Acta2*, consistent with a myofibroblast phenotype (Fig. 1b). To understand the potential functions of these uncharacterized populations we performed pathway analyses. We observed a robust enrichment of arachidonic acid metabolism genes in F3 (Cajal) cells and in the small population of F2d fibroblasts (Extended Data Fig. 1g, Supplementary Table 1), a pathway strongly associated with colorectal cancer⁶. Arachidonic acid is processed by cyclooxygenases to prostanoids, highly bioactive lipid mediators with a very short half-life and an autocrine or paracrine function in tissues⁶. In humans, pharmacological inhibition of cyclooxygenase-2 (Cox-2, also known as *Ptgs2*) prevents both hereditary and sporadic forms of colorectal cancer through an unknown cellular mechanism, but adverse cardiovascular effects currently impede its clinical application⁷. By fractionating normal human colonic tissues we found that expression of the Cox-2-encoding gene *PTGS2* is nearly undetectable in the epithelium but occurs predominantly in stromal cells; the same pattern as observed in the mouse intestine (Extended Data Fig. 1c, d). Our single-cell analyses showed that in the steady-state, mouse intestine *Ptgs2* is predominantly expressed in F3 (Cajal) cells and in the *Pdgfra*^{low}*Fgfr2*⁺*Vcam1*^{high} F2d fibroblasts (cluster 13) (Fig. 1c, Extended Data Fig. 1f). In humans, *PTGS2* is mainly expressed in *PDGFRA*^{low}*FGFR2*⁺*VCAM1*⁺ fibroblasts (cluster 8) and, to a lesser extent, in other fibroblast populations (Extended Data Fig. 1h).

Immunostaining of Cox-2 protein in the mouse intestine verified the presence of a Cox-2-expressing fibroblast population in the muscular layer, a location consistent with that of Cajal cells, and a second rare Cox-2-expressing fibroblast population located around part of the crypts, in close proximity to the stem-cell zone, suggestive of the F2d fibroblast cluster (Fig. 1d). We named these cells rare pericryptal *Ptgs2*-expressing fibroblasts (RPPFs). In agreement with their pericryptal location, RPPFs are marked by expression of the laminin subunit A1 (encoded by *Lama1*), a basement membrane protein detected specifically at the mesenchymal–epithelial interface, and also by expression of R-spondin 1 (encoded by *Rspo1*), a stem-cell niche factor detected mainly at the crypt base (Fig. 1b, Extended Data Fig. 2e, f).

Given the localization of RPPFs within the stem-cell microenvironment, we aimed to understand their role as a potential mesenchymal niche of tumour-initiating stem cells. We used fibroblast-specific *Col6-cre* mice, which target a substantial fraction of *Pdgfra*⁺ intestinal fibroblasts, including fibroblasts surrounding the crypts and Cox-2-expressing fibroblasts, as shown by lineage tracing, flow cytometry in reporter mice and quantitative PCR with reverse transcription (RT–qPCR) analyses in *Col6-Cre*⁺ cells separated by fluorescence-activated cell sorting (FACS) (Extended Data Fig. 3a–c). Specific ablation of Cox-2 in *Col6-cre*⁺ fibroblasts in *Col6-cre-Ptgs2*^{fl/fl} (*Ptgs2*^{ΔFibr}) mice was efficient and led to a significant reduction of *Ptgs2* expression levels in the whole tissue (Extended Data Fig. 3d, e), thereby confirming that fibroblasts are a predominant source of Cox-2 in the steady-state intestine.

To address the role of RPPFs in the mesenchymal niche in early tumour initiation, we used the *Apc*^{Min/+} mouse model in which autochthonous intestinal tumours are spontaneously formed by stem cells losing heterozygosity⁸. This model is highly relevant to human cancer, since somatic or germline mutations in the *APC* gene, a negative regulator of Wnt– β -catenin signalling, drive sporadic or hereditary forms, respectively, of intestinal neoplasia. Intestinal stem cell-specific *Apc* ablation is sufficient to drive tumorigenesis⁸. Notably, although adenoma formation in *Apc*-mutant mice is known to be Cox-2-dependent, the pathway by which this is mediated remains unknown^{9,10}. Thus, we generated *Apc*^{Min/+}*Ptgs2*^{ΔFibr} mice and studied adenoma formation. We found that specific ablation of *Ptgs2* in fibroblasts led to a strong reduction in the number of microadenomas formed in the small intestine at the early stage of 5 weeks (Fig. 2a) and significantly fewer macroscopic tumours formed at 5.5 months (Fig. 2b), along with a milder splenomegaly and a

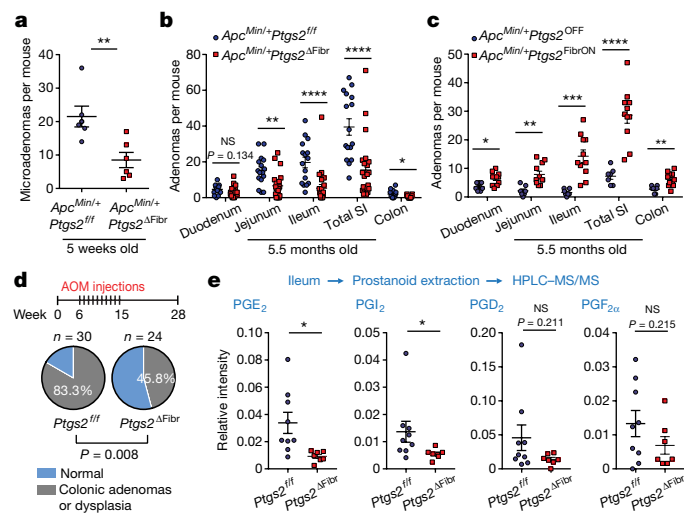


Fig. 2 | *Ptg2*-expressing fibroblasts drive tumour initiation by secreting PGE₂ in the crypt microenvironment. **a**, Number of β -catenin⁺ microadenomas in two sections of the small intestine (SI) of 5-week-old *Apc^{Min/+} Ptg2^{fl/fl}* (n = 6) and *Apc^{Min/+} Ptg2^{ΔFibr}* (n = 6) mice. Two-tailed t-test. **b**, Number of macroscopic adenomas in 5.5-month-old *Apc^{Min/+} Ptg2^{fl/fl}* (n = 16) and *Apc^{Min/+} Ptg2^{ΔFibr}* (n = 23) mice. Two-tailed Mann-Whitney test. **c**, Number of macroscopic adenomas in 5.5-month-old *Apc^{Min/+} Ptg2^{OFF}* (n = 7) mice in which *Ptg2* expression is blocked, and *Apc^{Min/+} Ptg2^{FibrON}* (n = 11) littermates in which *Ptg2* is exclusively expressed in fibroblasts (Extended Data Fig. 3i). Two-tailed Mann-Whitney test (duodenum), t-test (jejunum and colon), Welch's t-test (ileum and total small intestine). **d**, Incidence of dysplasia and microadenoma development in the colon of *Ptg2^{fl/fl}* (n = 30) and *Ptg2^{ΔFibr}* (n = 24) mice treated with 10 weekly injections of azoxymethane (AOM). Two-sided Fisher's exact test. **e**, HPLC-MS/MS analysis of prostanoids in the ileum of littermate *Ptg2^{fl/fl}* (n = 9) and *Ptg2^{ΔFibr}* (n = 7) mice. The relative intensity to the respective internal standard is shown. Two-tailed t-test (PGE₂ and PGF_{2α}) and Mann-Whitney test (PGI₂ and PGD₂). Data are mean ± s.e.m. NS, non-significant; *P < 0.05; **P < 0.01; ***P < 0.001; ****P < 0.0001.

significantly prolonged survival (Extended Data Fig. 3f, g). We observed no difference in tumour size in *Apc^{Min/+} Ptg2^{ΔFibr}* mice (Extended Data Fig. 3h), which showed that mesenchymal Cox-2 is necessary for tumour initiation but not for tumour growth. To understand how critical Cox-2 expression in the mesenchymal niche is for tumour initiation, compared with other cellular sources of prostanoids, we examined whether selective Cox-2 expression in fibroblasts is sufficient to drive tumour initiation in *Apc*-mutant mice. For this purpose, we genetically engineered mice in which a loxP-stop-loxP cassette was knocked into the *Ptg2* gene, thereby preventing its expression (*Ptg2^{OFF}*) unless excised by Cre-mediated recombination, which reactivates the gene (*Ptg2^{ON}*) (Extended Data Fig. 3i). By crossing these with *Col6-cre* mice, we generated mice in which *Ptg2* is expressed exclusively in fibroblasts (*Ptg2^{FibrON}*). We found that control *Apc^{Min/+} Ptg2^{OFF}* mice developed only a few intestinal tumours, consistently with the phenotype of *Ptg2*-knockout mice in this model⁹. By contrast, *Apc^{Min/+} Ptg2^{FibrON}* mice—in which *Ptg2* is expressed exclusively in fibroblasts—displayed robust tumorigenesis and developed, on average, 30 adenomas per mouse in the small intestine (Fig. 2c). Thus, Cox-2 in fibroblasts is both necessary and sufficient for tumour initiation in *Apc*-mutant mice. To further establish the role of Cox-2-expressing fibroblasts in controlling tumorigenesis, we used a model of sporadic colonic tumorigenesis, which is driven by random mutations caused by repeated injections of azoxymethane, a potent mutagenic agent. We found that *Ptg2^{ΔFibr}* mice displayed a significantly lower incidence of dysplasia and adenoma formation in the colon at 28 weeks of age (Fig. 2d), along with a reduced number—but not reduced size—of adenomas and dysplastic foci per

mouse (Extended Data Fig. 3j). These results from two different models demonstrated in vivo that fibroblasts utilize the Cox-2 pathway to provide a tumour initiation-conducive microenvironment for mutated stem cells in the intestine. Thus, we show that resident fibroblasts physiologically control tumorigenesis in *trans*.

To identify which Cox-2-mediated prostanoids are secreted by fibroblasts in the crypt microenvironment in vivo and drive tumour initiation, we performed lipidomic analyses by liquid chromatography–tandem mass spectrometry (HPLC–MS/MS) in the intestine of *Ptg2^{ΔFibr}* mice. We identified a significant reduction in the relative abundance of PGE₂ and prostaglandin I₂ (PGI₂) in the whole tissue (Fig. 2e), consistent with the expression of the respective synthases, *Ptges* and *Ptgis*, in fibroblasts (Extended Data Fig. 1g), and a trend for decreased prostaglandin D₂ (PGD₂) and prostaglandin F_{2α} (PGF_{2α}) abundance. Out of these prostaglandins, PGE₂ promotes sporadic tumour formation^{11,12} but its cellular source, its cellular target and the receptor through which it acts have remained unknown. These are important considerations for therapeutic targeting in light of the adverse effects of Cox-2 inhibition⁷. Since the estimated half-life of PGE₂ in vivo is less than 15 s (ref. ¹³), we hypothesized that PGE₂ secreted by RPPFs may diffuse through the thin (less than 1 μm) basal lamina matrix and reach the neighbouring mutant stem cells at a concentration sufficient to signal and drive tumour initiation.

To study the effect of PGE₂ on intestinal stem-cell function we cultured crypts in organoid growth media (OGM) supplemented daily with 16,16-dimethyl PGE₂ (dmPGE₂), a derivative of PGE₂ with a prolonged half-life. Notably, PGE₂ prevented the formation of budding organoids, leading instead to the development of spheroid-like structures which lack the typical crypt-villus architecture (Extended Data Fig. 4a). This morphology is associated with poor differentiation and increased stemness¹⁴. We functionally assessed stem-cell activity and found that PGE₂-driven spheroids contain more stem cells with full organoid-forming capacity (Extended Data Fig. 4b). PGE₂ signals through four receptors, EP1–EP4 (encoded by *Ptger1*–*Ptger4*, respectively), all of which are expressed in the mouse intestine (Extended Data Fig. 5a). We found that EP4 is the major PGE₂ receptor expressed in the mouse intestinal epithelium, in stem and progenitor cells and in the normal human colon (Extended Data Fig. 5b–d). We then generated intestinal epithelial cell-specific *Villin-cre-Ptger4^{fl/fl}* mice (*Ptger4^{ΔIEC}*) and found that unlike control crypts, crypts from these mice were able to form budding organoids even when dmPGE₂ was added to the OGM (Fig. 3a). On the basis of these results, we focused on the specific role of epithelial *Ptger4* in tumour initiation.

To model the mesenchymal niche of intestinal stem cells, we developed a 3D organotypic co-culture comprised of primary intestinal fibroblasts and fresh crypts growing in OGM. When co-cultured with fibroblasts, the majority of crypts developed into spheroids rather than organoids (Fig. 3b). Addition of *Ptger4* inhibitors or co-culture with *Ptger4^{ΔIEC}* crypts was sufficient to restore the growth of budding organoids (Fig. 3c, Extended Data Fig. 4c). Thus, organotypic cultures show that fibroblasts control stem-cell function via paracrine PGE₂–*Ptger4* signalling.

To understand the specific effects of fibroblast-derived PGE₂ on stem-cell function and differentiation we performed single-cell RNA-seq in crypt–fibroblast co-cultures in which *Ptger4* signalling was either active (*Ptger4*-ON) or inhibited (*Ptger4*-OFF) (Fig. 3d). We analysed 2,192 cells out of which 1,585 were epithelial (Extended Data Fig. 4d). By clustering and aligning these cells with signatures of known intestinal epithelial populations¹⁵ (Extended Data Fig. 4e), we observed a markedly different cellular composition between *Ptger4*-ON and *Ptger4*-OFF co-cultures (Fig. 3d, e). First, fibroblast-derived PGE₂–*Ptger4* signalling prevented the terminal differentiation of enterocytes but not that of goblet or Paneth cells (Fig. 3e, f). Second, it induced a substantial expansion of a non-cycling population displaying an intermediate transcriptional profile between stem and tuft cells (cluster 3) (Fig. 3f–h,

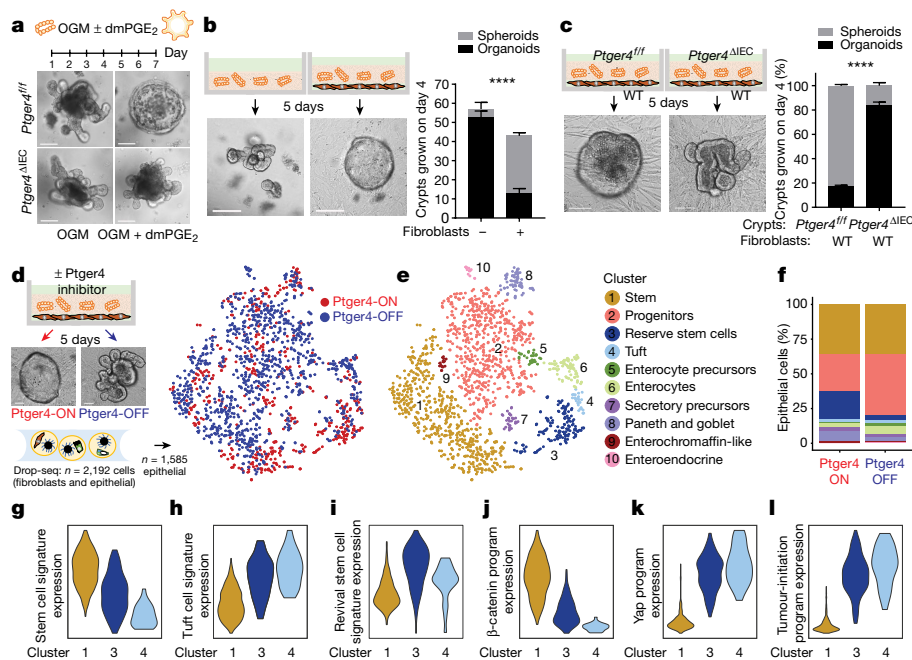


Fig. 3 | Fibroblast-derived PGE₂ drives the expansion of RSCs with a Yap-driven pro-tumorigenic program via receptor Ptger4. **a**, Indicative 3D cultures of crypts isolated from *Ptger4*^{fl/fl} and *Ptger4*^{ΔIEC} mice grown with OGM or OGM supplemented with 0.1 μM dmPGE₂. Scale bar, 100 μm. **b**, Normal crypts were cultured as 3D organoids in OGM (*n* = 2) or in co-cultures with primary mouse intestinal fibroblasts (*n* = 3). The absolute numbers of organoids or spheroids per 3D culture are shown. Results are representative of five independent experiments. Scale bar, 200 μm. **c**, Crypts isolated from *Ptger4*^{fl/fl} and *Ptger4*^{ΔIEC} mice were grown in co-cultures with wild-type (WT) primary mouse intestinal fibroblasts (*n* = 3 per genotype). The percentage of organoids and spheroids grown per 3D culture is shown. Results are representative of two

independent experiments. Scale bar, 100 μm. In **b**, **c**, data are mean ± s.e.m.; two-way ANOVA, *****P* < 0.0001. **d–l**, Single-cell RNA-seq of intestinal crypt-fibroblast co-cultures grown in OGM with 10 μM Ptger4 inhibitor (Ptger4-OFF) or DMSO (Ptger4-ON). Analyses are shown for 1,585 epithelial cells. **d**, t-SNE plot indicating epithelial cells from co-cultures with Ptger4-ON or Ptger4-OFF. **e**, t-SNE plot with clustering results. **f**, Proportion of each epithelial cluster among total epithelial cells in co-cultures with Ptger4-ON or Ptger4-OFF. **g–l**, Violin plots showing the entire range of metagene expression levels per single cell per cluster for the signatures/transcriptional programs of stem cells (**g**), tuft cells (**h**), RSCs (**i**), β-catenin (**j**), Yap (**k**) and early (non-tumour) *Apc*^{min/+} tumorigenesis (**l**).

Extended Data Fig. 4f). These cells express specific marker genes such as *Ly6a* (which encodes Sca-1), *Clu*, *Msln* and *Il1rn* (Extended Data Fig. 6b). *Clu* is a marker of revival stem cells¹⁶, a quiescent population that functions as reserve stem cells (RSCs)²⁴ and is induced upon irradiation damage in the intestinal epithelium. We found that the overall transcriptional program of cluster 3 cells is highly similar to that of RSCs (Fig. 3i, Extended Data Fig. 4g). Furthermore, *Ptger4* is expressed in RSCs and is strongly induced following irradiation damage in the regenerative intestinal epithelium (Extended Data Fig. 5e).

To understand the molecular link between PGE₂–Ptger4 and the RSC phenotype, we first tested whether PGE₂ activates the β-catenin pathway as reported in wound-associated epithelial cells¹⁷. However, we found a strong downregulation of the β-catenin transcriptional program in RSCs compared with cycling stem cells and no overall difference in this pathway between the Ptger4-ON and Ptger4-OFF conditions (Fig. 3j, Extended Data Fig. 4h, k). By contrast, we observed that Ptger4-ON spheroids overexpressed a set of genes reported to be targets of Yap¹⁸ as well as overall overexpression of a Yap transcriptional program¹⁸ (Extended Data Figs. 4k, 6a, b). This effect was validated in independent PGE₂-driven spheroid cultures and confirmed genetically to be mediated by Ptger4 (Extended Data Fig. 6c–e). Yap is a transcriptional effector of Hippo signalling, which is involved in stemness, organ size control, tissue homeostasis, regeneration and cancer¹⁹ and is key for RSC function¹⁶. Yap is also indispensable for tumorigenesis driven by *Apc*-deficient stem cells^{18,20}. Indeed, we found that a signature of early *Apc*^{min/+} tumorigenesis correlated with the Yap program (Extended Data Fig. 6f) and both were predominantly expressed in RSCs (Fig. 3k, l, Extended Data Fig. 4i, j).

To directly examine whether Yap mediates fibroblast-driven expansion of RSCs, we isolated crypts from intestinal epithelial cell-specific

Villin-cre-Yap^{fl/fl} mice (*Yap*^{ΔIEC}) and co-cultured them with fibroblasts in OGM. Although *Yap*^{ΔIEC} crypts require epiregulin supplementation to grow¹⁸, fibroblasts supported their growth in the absence of exogenous epiregulin (Extended Data Fig. 7a). Notably, unlike *Yap*^{fl/fl} crypts, *Yap*^{ΔIEC} crypts did not form spheroids in these co-cultures, but instead retained their crypt morphology (Extended Data Fig. 7a). Flow cytometry analyses for the RSC marker Sca-1 showed a robust expansion of Sca-1⁺ cells in co-cultures of fibroblasts with *Yap*^{fl/fl} crypts which was prevented in co-cultures with *Yap*^{ΔIEC} crypts (Extended Data Fig. 7b). Collectively, these analyses revealed that fibroblast-derived PGE₂ drives the expansion of an RSC population with a regenerative/tumorigenic program via Ptger4 and Yap.

Next, we examined how PGE₂ activates a Yap transcriptional program in crypts. In contrast to an earlier study using a cancer cell line²¹, stimulation of intestinal organoids with PGE₂ did not induce *Yap1* expression at the RNA or at the protein level (Extended Data Fig. 6g, h). In addition, day 3 PGE₂-driven spheroids showed no difference in total Yap protein expression (Extended Data Fig. 6g). Since G-protein-coupled receptor (GPCR) signalling has been suggested to either activate or inhibit the Hippo–Yap signalling pathway²², and Ptger4 is a GPCR, we hypothesized that PGE₂ may activate Yap in the intestinal epithelium by inhibiting its regulation by the Hippo pathway. Indeed, when we stimulated wild-type organoids with dmPGE₂, we observed Yap dephosphorylation at Ser127 within 30–60 min (Extended Data Fig. 6i), suggestive of inhibition of Hippo activity and activation of Yap. This effect was mediated by Ptger4 (Fig. 4a). Furthermore, stimulation of wild-type organoids with dmPGE₂ led to nuclear translocation of Yap within 30–60 min (Fig. 4b) and transcriptional activation of Yap target genes (Extended Data Fig. 6h), which was prevented by a Ptger4 inhibitor (Fig. 4c). PGE₂-stimulation experiments with *Yap*^{ΔIEC} organoids or wild-type organoids

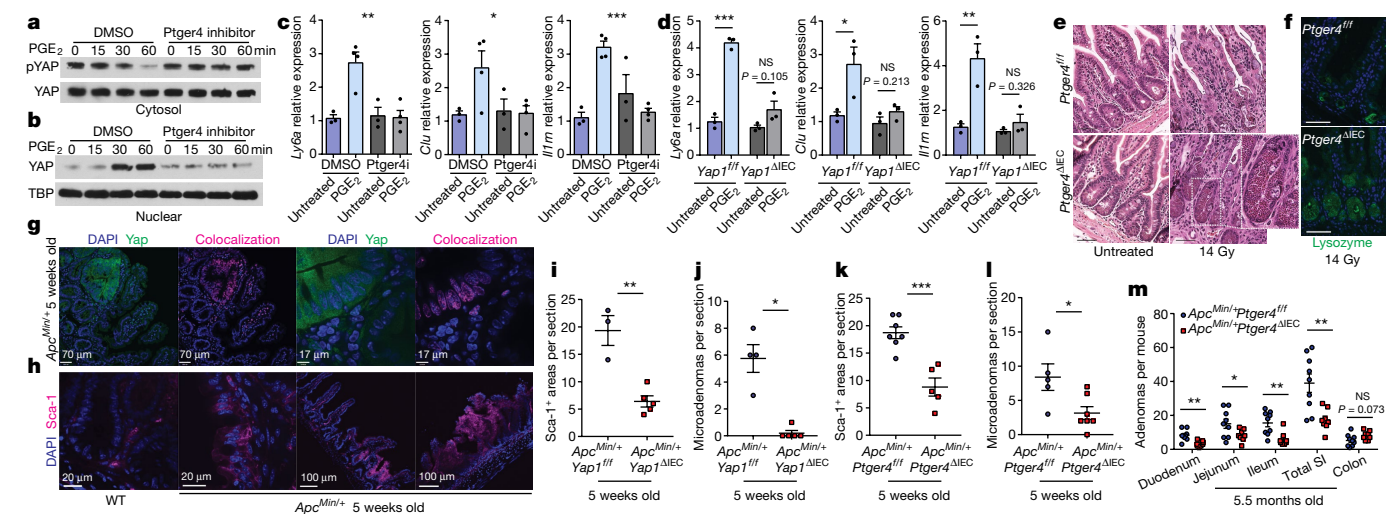


Fig. 4 | Epithelial *Ptger4* induces Yap nuclear translocation, mediates RSC mobilization and drives tumour initiation. **a, b**, Wild-type organoids pretreated with or without 10 μ M *Ptger4* inhibitor were stimulated with 0.1 μ M dmPGE₂. Western blots for phosphorylated Yap Ser127 (pYap) and total Yap in cytoplasmic lysates (**a**), Yap and TBP in nuclear lysates (**b**). Results are representative of two experiments. **c, d**, Relative expression of Yap target genes¹⁸. **c**, Wild-type organoids treated with 10 μ M *Ptger4* inhibitor and 0.1 μ M dmPGE₂ for 13 h. $n = 3$ to 4 cultures per condition. One-way ANOVA. **d**, *Yap1* ^{Δ IEC} and *Yap1*^{fl/fl} organoids treated with 0.1 μ M dmPGE₂ for 13 h. Three cultures per genotype per condition. Two-tailed *t*-test. **e, f**, *Ptger4*^{fl/fl} ($n = 3$) and *Ptger4* ^{Δ IEC} ($n = 3$) mice received 14 Gy of abdominal irradiation. On day 3, the ileum was analysed by haematoxylin and eosin staining (**e**) and immunostaining for lysozyme (**f**). Results are representative of three independent experiments. Scale bars, 50 μ m. **g**, Immunostaining for Yap in the small intestine of five-week-

old wild-type and *Apc*^{Min/+} mice. Nuclear Yap is evaluated on the basis of colocalization with DAPI. Results are indicative of at least eight microadenomas. **h**, Immunostaining for Sca-1 in the small intestine of five-week-old wild-type and *Apc*^{Min/+} mice. Results are representative of two independent experiments. **i, k**, Quantification of areas of the small intestine with expansion of Sca-1⁺ epithelial cells in *Apc*^{Min/+}*Yap1*^{fl/fl} ($n = 3$) and *Apc*^{Min/+}*Yap1* ^{Δ IEC} ($n = 5$) mice (**i**), and *Apc*^{Min/+}*Ptger4*^{fl/fl} ($n = 7$), *Apc*^{Min/+}*Ptger4* ^{Δ IEC} ($n = 5$) mice (**k**). Two-tailed *t*-test. **j, l**, Number of BrdU⁺ microadenomas per small-intestinal section of *Apc*^{Min/+}*Yap1*^{fl/fl} ($n = 4$) and *Apc*^{Min/+}*Yap1* ^{Δ IEC} ($n = 5$) mice (**j**), and *Apc*^{Min/+}*Ptger4*^{fl/fl} ($n = 5$) and *Apc*^{Min/+}*Ptger4* ^{Δ IEC} ($n = 7$) mice (**l**). Two-tailed Mann–Whitney test (**j**); two-tailed *t*-test (**l**). **m**, Number of macroscopic adenomas in 5.5-month-old *Apc*^{Min/+}*Ptger4*^{fl/fl} ($n = 9$) and *Apc*^{Min/+}*Ptger4* ^{Δ IEC} ($n = 8$) mice. Two-tailed *t*-test (duodenum and colon); Welch's *t*-test (jejunum and total small intestine); Mann–Whitney test (ileum). Data are mean \pm s.e.m.

treated with verteporfin (an inhibitor of the Yap–Tead interaction²³) demonstrated that the activation of these genes is mediated by Yap (Fig. 4d, Extended Data Fig. 6j). These results establish that PGE₂–*Ptger4* signalling inhibits Hippo activity and, consequently, leads to Yap nuclear translocation and induction of a Yap–Tead-dependent gene-expression program in the intestinal crypt.

Epithelial ablation of *Ptger4* in *Ptger4* ^{Δ IEC} mice is efficient but does not affect stem-cell function and epithelial lineage differentiation in the steady-state intestine, as assessed by single-cell RNA-seq, 5-bromo-2'-deoxyuridine (BrdU)-incorporation experiments and immunostaining for population markers (Extended Data Fig. 8a–f, i). Lineage tracing of *Ptger4*-deficient *Lgr5*⁺ stem cells confirmed that *Ptger4* is dispensable for the function of normal stem cells at the steady state (Extended Data Fig. 8h). Similar results were obtained in *Ptgs2* ^{Δ Fib} mice (Extended Data Fig. 3k). To assess the role of PGE₂–*Ptger4*–Yap in stem-cell reprogramming in vivo, we employed an abdominal-irradiation-induced injury model in the context of which slow-cycling RSC populations are mobilized and mediate the epithelial regenerative response²⁴. In the absence of Yap, although no phenotype is observed at the steady state, the epithelial response to irradiation is perturbed, causing increased Paneth cell differentiation¹⁸. We found that exposure of *Ptger4* ^{Δ IEC} mice to 14 Gy of abdominal irradiation led to a pronounced expansion of Paneth cells three days after irradiation (Fig. 4e, f), thereby phenocopying the effect of Yap deficiency. These results functionally validate the critical function of *Ptger4* for RSC mobilization in vivo.

In early tumour initiation, we found that Yap displays an increased nuclear localization in microadenomas of five-week-old *Apc*^{Min/+} mice (Fig. 4g). Furthermore, we observed that Sca-1, a Yap target gene and RSC marker, is detected in the mesenchyme but not in the epithelium in the steady-state in wild-type mice. By contrast, however, in the intestine of five-week-old *Apc*^{Min/+} mice, we observed areas of

the epithelium in which Sca-1⁺ epithelial cells were detected as local expansions, whereas Sca-1⁺ epithelial cells were more widespread in microadenomas (Fig. 4h). Similar expansion of Sca-1⁺ cells has been described in regenerative contexts such as the response to irradiation and to helminth infection²⁵. Given these data, we examined the role of Yap in the expansion of Sca-1⁺ cells and in tumour initiation. In five-week-old *Apc*^{Min/+}*Yap1* ^{Δ IEC} mice, we found a markedly decreased number of Sca-1⁺ areas in the epithelium compared with littermate *Apc*^{Min/+}*Yap1*^{fl/fl} controls, as well as an almost completely abrogated formation of microadenomas (Fig. 4i, j). These results show that in early tumour initiation, Yap translocates to the nucleus and drives the expansion of Sca-1⁺ cells and the formation of microadenomas. Nuclear localization of Yap and epithelial Sca-1 expression were also observed in developed tumours in five-month-old *Apc*^{Min/+} mice (Extended Data Fig. 9a, b). Moreover, mice treated with repeated azoxymethane injections displayed nuclear localization of Yap in tumours and overexpression of the Yap target gene *Clu* (Extended Data Fig. 9c, d).

To address whether fibroblast-derived PGE₂ activates the Yap program and drives tumour initiation via epithelial *Ptger4*, we generated *Apc*^{Min/+}*Ptger4* ^{Δ IEC} mice. Given the crucial role of Yap in tumour initiation, we first examined whether *Ptger4* mediates the activation of Yap target genes in these mice. We found that at five weeks of age, *Apc*^{Min/+} mice displayed an increased expression of Yap target genes, but not of *Yap1* itself, compared with normal controls; however, in *Apc*^{Min/+}*Ptger4* ^{Δ IEC} littermates, this upregulation of the same Yap targets was abrogated (Extended Data Fig. 9e). Most notably, five-week-old *Apc*^{Min/+}*Ptger4* ^{Δ IEC} mice displayed an attenuated expansion of Sca-1⁺ cells in the epithelium compared with their *Apc*^{Min/+}*Ptger4*^{fl/fl} littermates (Fig. 4k). Moreover, *Apc*^{Min/+}*Ptger4* ^{Δ IEC} mice developed fewer microadenomas at five weeks of age (Fig. 4l). They also displayed a strong reduction in the number of macroscopic tumours formed at 5.5 months (Fig. 4m), a significantly

alleviated splenomegaly (Extended Data Fig. 9f) and—consistent with a role of *Ptger4* in tumour initiation rather than tumour growth—no difference in the size of the tumours formed (Extended Data Fig. 9g). These results provide definitive genetic evidence that epithelial *Ptger4* is the receptor that mediates the tumorigenic effect of PGE_2 and explain the role of RPPFs as paracrine drivers of tumour initiation.

On the basis of these results, we addressed the role of PGE_2 –*PTGER4* in the human intestinal stem-cell niche. By isolating crypts from normal parts of the colon from three patients and culturing them in OGM we found that PGE_2 drives the formation of spheroid-like structures. This effect was fully prevented by treatment with a *PTGER4* inhibitor, thus confirming that PGE_2 –*PTGER4* also controls stem-cell function in the human colonic crypt (Extended Data Fig. 10a). Furthermore, we performed immunostaining for YAP in tissues from 16 patients, including individuals with sporadic adenomas and adenocarcinomas, familial adenomatous polyposis, Lynch syndrome and cancer associated with inflammatory bowel disease (Supplementary Table 3). We observed that YAP displayed a nuclear localization in tumours but not in the neighbouring normal areas of the tissue in all these samples (Extended Data Fig. 10b), supporting its role in tumorigenesis. Of note, both *PTGER4* and *YAP1* genetic loci were recently identified to be genetically associated with colorectal cancer risk in genome-wide associations studies²⁶, further underlining the relevance of this pathway to human disease.

The results of this study show that PGE_2 -secreting RPPFs provide a micro-niche favouring the activation of the pro-tumorigenic Yap program in neighbouring stem cells, thereby driving tumorigenesis in the presence of mutations (Extended Data Fig. 10c). This work establishes in vivo that the formation of intestinal tumours requires the paracrine interaction of mutated stem cells with their native mesenchymal microenvironment.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2166-3>.

1. Vermeulen, L. & Snippert, H. J. Stem cell dynamics in homeostasis and cancer of the intestine. *Nat. Rev. Cancer* **14**, 468–480 (2014).
2. Powell, D. W., Pinchuk, I. V., Saada, J. I., Chen, X. & Mifflin, R. C. Mesenchymal cells of the intestinal lamina propria. *Annu. Rev. Physiol.* **73**, 213–237 (2011).
3. Kinchen, J. et al. Structural remodeling of the human colonic mesenchyme in inflammatory bowel disease. *Cell* **175**, 372–386 (2018).

4. Hamilton, T. G., Klinghoffer, R. A., Corrin, P. D. & Soriano, P. Evolutionary divergence of platelet-derived growth factor alpha receptor signaling mechanisms. *Mol. Cell. Biol.* **23**, 4013–4025 (2003).
5. Molotkov, A., Mazot, P., Brewer, J. R., Cinalli, R. M. & Soriano, P. Distinct requirements for FGFR1 and FGFR2 in primitive endoderm development and exit from pluripotency. *Dev. Cell.* **41**, 511–526 (2017).
6. Smyth, E. M., Grosser, T., Wang, M., Yu, Y. & FitzGerald, G. A. Prostanoids in health and disease. *J. Lipid Res.* **50** (Suppl.), S423–S428 (2009).
7. Wang, D. & DuBois, R. N. The role of anti-inflammatory drugs in colorectal cancer. *Annu. Rev. Med.* **64**, 131–144 (2013).
8. Barker, N. et al. Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457**, 608–611 (2009).
9. Chulada, P. C. et al. Genetic disruption of *Ptgs-1*, as well as *Ptgs-2*, reduces intestinal tumorigenesis in *Min* mice. *Cancer Res.* **60**, 4705–4708 (2000).
10. Cherukuri, D. P. et al. Targeted *Cox2* gene deletion in intestinal epithelial cells decreases tumorigenesis in female, but not male, *Apc^{Min/+}* mice. *Mol. Oncol.* **8**, 169–177 (2014).
11. Xia, D., Wang, D., Kim, S. H., Katoh, H. & DuBois, R. N. Prostaglandin E_2 promotes intestinal tumor growth via DNA methylation. *Nat. Med.* **18**, 224–226 (2012).
12. Wang, D., Fu, L., Sun, H., Guo, L. & DuBois, R. N. Prostaglandin E_2 promotes colorectal cancer stem cell expansion and metastasis in mice. *Gastroenterology* **149**, 1884–1895 (2015).
13. Bygdeman, M. Pharmacokinetics of prostaglandins. *Best Pract. Res. Clin. Obstet. Gynaecol.* **17**, 707–716 (2003).
14. Mustata, R. C. et al. Identification of *Lgr5*-independent spheroid-generating progenitors of the mouse fetal intestinal epithelium. *Cell Rep.* **5**, 421–432 (2013).
15. Haber, A. L. et al. A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333–339 (2017).
16. Ayyaz, A. et al. Single-cell transcriptomes of the regenerating intestine reveal a revival stem cell. *Nature* **569**, 121–125 (2019).
17. Miyoshi, H. et al. Prostaglandin E_2 promotes intestinal repair through an adaptive cellular response of the epithelium. *EMBO J.* **36**, 5–24 (2017).
18. Gregorieff, A., Liu, Y., Inanlou, M. R., Khomchuk, Y. & Wraana, J. L. Yap-dependent reprogramming of *Lgr5⁺* stem cells drives intestinal regeneration and cancer. *Nature* **526**, 715–718 (2015).
19. Hong, A. W., Meng, Z. & Guan, K. L. The Hippo pathway in intestinal regeneration and disease. *Nat. Rev. Gastroenterol. Hepatol.* **13**, 324–337 (2016).
20. Cai, J., Maitra, A., Anders, R. A., Taketo, M. M. & Pan, D. β -Catenin destruction complex-independent regulation of Hippo–YAP signaling by APC in intestinal tumorigenesis. *Genes Dev.* **29**, 1493–1506 (2015).
21. Kim, H. B. et al. Prostaglandin E_2 activates YAP and a positive-signaling loop to promote colon regeneration after colitis but also carcinogenesis in mice. *Gastroenterology* **152**, 616–630 (2017).
22. Yu, F. X. et al. Regulation of the Hippo–YAP pathway by G-protein-coupled receptor signaling. *Cell* **150**, 780–791 (2012).
23. Liu-Chittenden, Y. et al. Genetic and pharmacological disruption of the TEAD–YAP complex suppresses the oncogenic activity of YAP. *Genes Dev.* **26**, 1300–1305 (2012).
24. Mills, J. C. & Sansom, O. J. Reserve stem cells: differentiated cells reprogram to fuel repair, metaplasia, and neoplasia in the adult gastrointestinal tract. *Sci. Signal.* **8**, re8 (2015).
25. Nusse, Y. M. et al. Parasitic helminths induce fetal-like reversion in the intestinal stem cell niche. *Nature* **559**, 109–113 (2018).
26. Huyghe, J. R. et al. Discovery of common and rare genetic risk variants for colorectal cancer. *Nat. Genet.* **51**, 76–87 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Article

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment unless otherwise stated in the Methods.

Mice

Ptgs2^{fl/fl} mice²⁷ crossed with *Col6-cre* mice²⁸, *Rosa26^{mT/mG}* mice²⁹ and *Apc^{Min/+}* mice³⁰ were bred in the animal facilities of the BSRC 'Alexander Fleming' under specific pathogen-free conditions. *Ptger4^{fl/fl}* mice³¹ and *Yap^{fl/fl}* mice³² crossed with *Villin-cre* mice³³ and *Apc^{Min/+}* mice³⁰, wild-type mice used for organoid experiments, *Lgr5-eGFP-IRES-creERT2* mice³⁴, *Pdgfra^{eGFP/+}* mice⁴, as well as *Col6-cre* mice²⁸ crossed with *Rosa26^{tdTomato/+}* mice (Ai14)³⁵ and *Ptgs2* Lox-Stop-Lox-knockin mice (*Ptgs2^{LSL}*); the generation of this mouse strain is described below) were bred in the facilities of the Yale Animal Resources Center. All these mice were maintained on a C57BL/6J genetic background. *Fgfr2-T2A-H2B-mCherry* mice⁵ were maintained in the facilities of the Icahn School of Medicine at Mount Sinai on the 129S4 genetic background. Mice were housed in standard cages, on a 12:12 h day:night cycle and were fed a standard rodent chow. Mice were used for experiments at 8–12 weeks of age unless otherwise indicated. For all experiments, littermate, co-housed and sex-matched mice were used. Both male and female mice were used for experiments. No mice were excluded from the analyses performed. End points used for mice developing tumours were changes in activity or mobility, abnormal posture, decreased food and/or water intake and decreased body temperature. Experiments in BSRC 'Alexander Fleming' were approved by the Institutional Committee of Protocol Evaluation in conjunction with the Veterinary Service Management of the Hellenic Republic Prefecture of Attika according to all current European and national legislation. All animal experimentation at Yale was performed in compliance with Yale Institutional Animal Care and Use Committee protocols.

Generation of *Ptgs2* Lox-stop-Lox-knockin mice

Ptgs2 flox-stop-knockin mice were generated by the University of California, Davis Mouse Biology Program services. JMB8 (C57BL/6N) embryonic stem (ES) cells were targeted with a vector containing a diphtheria toxin A (PGK-DTA) cassette, a 4 kb 5' arm of homology, two loxP sites within intron 3 of the *Ptgs2* gene flanking a STOP cassette sequence (derived from Addgene plasmid 11584), and a frt-flanked PKG-neomycin cassette and a 5.1 kb 3' arm of homology. The PKG-neomycin element enabled positive selection in ES cells, while the DTA element enabled negative selection in ES cells. Mice bearing the targeted lox-stop-frt-PKG-neomycin-frt-lox *Ptgs2* allele in the germline were crossed with the B6N(B6J)-Tg(CAG-Flpo)1Afst/Mmucd transgenic mouse (Mutant Mouse Resource and Research Center MMRC_036512-UCD) and the PKG-neomycin cassette was removed and mice bearing a lox-stop-frt-lox *Ptgs2* allele (*Ptgs2^{LSL}*) were obtained (Extended Data Fig. 3i).

Human study participants

Fresh human colon tissue was obtained from the Yale Pathology Archives on the basis of Yale Human Investigation Committee protocols no. 0304025173, which allows retrieval of tissue from surgical pathology that was consented or has been approved for use with waiver of consent. The data were analysed anonymously from preexisting patient databases and are thus exempt from consent by the human studies committee. Patient characteristics (sex, age, diagnosis) are described in the Supplementary Table 3. All tissue segments were obtained from the uninvolved surgical margins of colon resections. The specific part of the colon resected is indicated in the Supplementary Table 3. Ischaemic time of all samples ranged from 1 h to 3 h. All collected samples were kept on ice-cold RPMI medium before processing.

Formalin-fixed paraffin-embedded colorectal tumour tissue was obtained from the Yale Pathology Archives. The data were analysed anonymously from preexisting patient databases and hence exempt from consent by the human studies committee. Patient characteristics (sex, age, diagnosis) are described in the Supplementary Table 3.

Isolation of human intestinal epithelial cells and stromal cells

For the isolation of intestinal epithelial and stromal cells the tissue was cut into 0.5 cm pieces and incubated five times in HBSS containing 0.5 mM EDTA and 1 mM DTT for 15 min, at 4 °C on a rocker. Epithelial cells were released by vigorous shaking and passed through a 70-µm strainer, washed and used for RNA isolation. For stromal cell isolation the tissue pieces were incubated in DMEM containing 10% FBS, 300 U ml⁻¹ Collagenase XI (Sigma, C7657), 0.1 mg ml⁻¹ Dispase II (Sigma, D4693) and 50 U ml⁻¹ DNase II Type V (Sigma, D8764) for 1 h, at 37 °C, 200 rpm. Cells released after vigorous shaking were passed through a 70-µm strainer, treated with ammonium-chloride-potassium red-blood-cell-lysing buffer, washed with 2% sorbitol and then used for RNA isolation.

Human colonic organoid culture

For the isolation of human colonic crypts the tissue was cut into 0.5-cm pieces and incubated six times in PBS containing 5 mM EDTA and 1 mM DTT for 10 min, at 4 °C on a rocker. Epithelial cells and whole crypts were released by vigorous shaking. The fractions enriched for crypts were further processed. Crypts were washed by centrifugation at 100g, 50g and 30g and then used for organoid development in domes made by Matrigel (Corning, 356231) and IntestiCult Organoid Growth Medium (Human) (Stem Cell Technologies, 06010) according to the manufacturer's guidelines. When indicated, 16,16-dimethyl PGE₂ (Cayman, 14750) dissolved in ethanol was added daily at a final concentration of 0.1 µM. Ethanol was used as a vehicle control for the untreated organoids. The ONO-AE3-208 *Ptger4* (EP4) inhibitor (Cayman, 14522) dissolved in DMSO was added at a final concentration of 10 µM 1 h before stimulation and DMSO was used as a vehicle control.

Isolation of mouse intestinal epithelial cells and mesenchymal cells

The intestine was dissected, flushed, opened longitudinally and then cut into 1 cm pieces. The tissues were incubated in HBSS containing 1 mM EDTA, 1 mM DTT, 0.2% FBS, 4–5 times, 10 min each, at 37 °C, 200 rpm. Epithelial cells were released by vigorous shaking, passed through a 70 µm strainer, washed and immediately lysed for RNA isolation. After epithelial cell removal, the remaining stromal part of the intestine was lysed for RNA isolation. For Drop-seq analysis or for FACS-sorting of mesenchymal cells the tissues were processed as above and then incubated in DMEM 10% FBS containing Collagenase XI (300 units/ml, Sigma, C7657), Dispase II (0.1 mg/ml, Sigma, D4693) and DNase II Type V (50 units/ml, Sigma, D8764) for 1 h, at 37 °C, 200 rpm. Cells released after vigorous shaking were passed through a 70 µm strainer and washed with 2% sorbitol. Such cell preparations were directly processed by Drop-seq or by flow cytometry as described below.

Mouse intestinal organoid culture and fibroblast/crypt organotypic co-culture

Crypts were isolated from the last three fourths of the small intestine. The intestine was flushed, cut longitudinally and the villi were scraped off with a glass coverslip. The tissue was then cut into 0.5 cm pieces which were incubated in PBS containing 5 mM EDTA, 0.2% FBS for 30 min at 4 °C on a rocker. Crypts were released by vigorous shaking and were passed through a 70 µm strainer. Six fractions were obtained after vigorous shaking and the ones enriched for crypts were further processed. Crypts were washed by centrifugation at 200g, 100g and 50g and then used for organoid development in domes made by Matrigel

(Corning, 356231) and IntestiCult Organoid Growth Medium (Stem Cell Technologies, 06005) according to manufacturer's guidelines. When indicated, dmPGE₂ (Cayman, 14750) dissolved in ethanol was added daily at a final concentration of 0.1 μ M. Ethanol was used as a vehicle control for the untreated organoids. Crypts isolated from *Yap1*^{ΔEC} mice were cultured in IntestiCult Organoid Growth Medium (Stem Cell Technologies, 06005) supplemented with 0.5 μ g ml⁻¹ recombinant mouse epiregulin (RnD1068-EP-050) or co-cultured with intestinal fibroblasts in OGM without epiregulin.

For the assessment of stem-cell activity, organoids or spheroids were dissociated into single cells by incubation at 37 °C in 0.25% trypsin-EDTA solution (Gibco, 25200056) diluted 1:1 with DMEM without serum. Numbers of live cells were counted after staining with trypan blue. In each experiment, the same number of live single cells per condition ($n = 3,000$ – $11,000$) were cultured in domes made by Matrigel (Corning, 356231) and OGM.

Intestinal organoids were stimulated with dmPGE₂ at a final concentration of 0.1 μ M. Ethanol was used as a vehicle control. The ONO-AE3-208 Ptger4 (EP4) inhibitor (Cayman, 14522) dissolved in DMSO was added at a final concentration of 10 μ M 1 h before stimulation. Verteporfin (Cayman, 17334) dissolved in DMSO was added at a final concentration of 1 μ M 1 h before stimulation. DMSO was used as a vehicle control for ONO-AE3-208 and Verteporfin.

Fibroblasts were isolated from the small intestine of mice. The intestine was dissected, flushed, opened longitudinally and then cut into 1-cm pieces. The tissues were incubated in HBSS containing 1 mM EDTA, 1 mM DTT and 0.2% FBS 4–5 times for 10 min each, at 37 °C, 200 rpm. Epithelial cells were released by vigorous shaking. Then, the tissues were incubated in DMEM 10% FBS containing Collagenase XI (300 U ml⁻¹, Sigma, C7657), Dispase II (0.1 mg ml⁻¹, Sigma, D4693) and DNase II Type V (50 U ml⁻¹, Sigma, D8764) for 1 h, at 37 °C, 200 rpm. Cells released after vigorous shaking were passed through a 70- μ m strainer, washed and cultured in DMEM with 10% FBS. For co-culture experiments 2×10^4 fibroblasts were seeded in 48-well plates overnight. Freshly isolated crypts ($n = 500$) were suspended in 1:1 Matrigel (Corning, 356231) and OGM and added as an overlay on the fibroblasts. Crypts and fibroblasts were co-cultured with OGM. When indicated, the ONO-AE3-208 Ptger4 (EP4) inhibitor dissolved in DMSO was added to the co-cultures every second day at a final concentration of 10 μ M. DMSO was used as a vehicle control for the untreated co-cultures.

Quantitative real-time PCR

RNA was isolated with the TRIzol reagent (Thermo Fisher, 15596026) followed by DNase I treatment (Roche, 04716728001) or with the QIAGEN RNA isolation RNeasy plus Mini Kit (QIAGEN, 74134) according to the manufacturer's instructions. Reverse transcription was performed with the Maxima H Minus Reverse Transcriptase (Thermo Fisher, EP0751). RT-qPCR analyses were performed using iTaq Fast SYBR Green Supermix (Bio-Rad, 1725100) and a CFX96 Touch Real-Time PCR Detection System (Bio-Rad). Data were acquired and analysed with the CFX Manager software (Bio-Rad). Gene expression relative to a control sample was calculated with the RelQuant software (Bio-Rad Laboratories) by normalizing to *B2m* expression. Where indicated, relative expression (RE) to *B2m* was calculated as $RE = 2^{-\Delta\Delta C_t}$. Primers used for human were *B2M*-F: ATGAGTATGCCTGCCGTG TG, *B2M*-R: CCAATGCGGCATCTTCAAAC, *PTGS2*-F: TGTGAAAAGTAG TTCTGGG, *PTGS2*-R: AAGCAGGCTAATACTGATAGG. Primers for mouse were *B2m*-F: TTCTGGTGCTGTCTCACTGA, *B2m*-R: CAGTATGTTCCGCTT CCCATTC, *Ptgs2*: QT00165347 (QIAGEN) and *Ptgs2*-F: TCCAACCTCT CCTACTACACCAG, *Ptgs2*-R: GGGTCAGGATGAAGTCTCTC, *Ptger1*-F: AAGTTTTGGATTCACTTCCC, *Ptger1*-R: GAAGGTGTTGAGATTC TTGG, *Ptger2*-F: CTGCGCTTTCACAATCTTTG, *Ptger2*-R: ACCCA AGGGTCAATTATAGAG, *Ptger3*-F: CGCCGCTATTGATAATGATG, *Ptger3*-R: TTCTTAGCAGCAGATAAACC, *Ptger4*-F: GTGCGGAGATCCA GATGGTC, *Ptger4*-R: TCACCACGTTTGCTGATATAAC, *Ly6a*-F: GAAAG

AGCTCAGGGACTGGAGTGTT, *Ly6a*-R: TTAGGAGGGCAGATGGGTAA GCAA, *Clu*-F: GCTGCTGATCTGGGACAATG, *Clu*-R: ACCTACTCCCTTGAG TGGACA, *Il1rn*-F: GCTCATTTGCTGGGACTTACAA, *Il1rn*-R: CCAGACTTGG CACAAGACAGG, *Cxcl16*-F: CCTTGTCTCTTGCCTTCTTCC, *Cxcl16*-R: TCCAAAGTACCCTGCCGTATC, *Msln*-F: CTTAGTCTTGGGTGGATA, *Msln*-R: TCTTCTGTCTTACAGCCA, *Yap1*-F: GATGCTCAGGAATTGAGAAC and *Yap1*-R: CTGTATCCATTTCATCCACAC.

Western blot

Total protein was extracted with RIPA lysis buffer. Nuclear and cytoplasmic fractions were extracted with NE-PER Nuclear and Cytoplasmic Extraction Reagents (Thermo Fisher, 78833). Protease inhibitors (Thermo Fisher, 87786) and phosphatase inhibitors (Thermo Fisher, 78420) were added. Antibodies against pYAP (Ser127) (D9W21) (Cell Signaling, 13008), YAP (D8H1X) (Cell Signaling, 14074) and TBP (D5C9H) (Cell Signaling, 44059) were used at a 1:1,000 dilution in 5% BSA, overnight. Antibodies against β -actin (clone C4, Santa Cruz sc-47778) were used at a 1:2,000 dilution in 5% BSA for 2 h at room temperature.

Immunofluorescence and imaging

Two-photon microscopy was performed with a LaVision TriM Scope II (LaVision Biotech) microscope equipped with a Chameleon Vision II (Coherent) two-photon laser in the In Vivo Imaging Facility of Yale School of Medicine. Fresh whole-mount specimens of the small intestine and the colon of a *Pdgfra*^{eGFP/+} mouse⁴ were prepared and analysed immediately after mice were euthanized.

Confocal imaging was performed with a Nikon-Ti microscope combined with UltraVox spinning disk (PerkinElmer) and data were analysed by using the Volocity software (PerkinElmer). Rosa26–tdTomato, *Pdgfra*–eGFP and *Fgfr2*–mCherry were detected by direct fluorescence. The tissues were dissected, fixed in 4% paraformaldehyde for 4 h at 4 °C, followed by incubation in 30% sucrose/PBS overnight at 4 °C. Tissue samples were frozen in OCT (Tissue-Tek) on dry ice and kept at –80 °C until sectioning. Sections 10 μ m thick were prepared with a cryostat (Leica). After washing with PBS, sections were mounted with Fluoroshield histology medium containing DAPI (Sigma, F6057). For *Lgr5*–eGFP and vimentin immunostaining, the terminal ileum of a *Lgr5*–eGFP–IRES–creERT2 mouse³⁴, was dissected, frozen and processed as above. The staining was performed with an Alexa Fluor 488-conjugated rabbit polyclonal GFP antibody at 1:200 (Thermo Fisher, A-21311) and an Alexa Fluor 647-conjugated rabbit monoclonal vimentin antibody at 1:200 (Cell Signaling, 9856), overnight at 4 °C. Sca-1 immunostaining was performed in frozen tissue sections processed as above and stained with a rat monoclonal Alexa Fluor 647 antibody (E13-161.7, Biolegend 122517, 1:400) overnight at 4 °C. The number of Sca-1⁺ areas was quantified in sections of small intestine prepared with the Swiss-roll technique and an EVOS FL Auto 2 Imaging System (Thermo). The evaluation was blinded to mouse genotype.

Formalin-fixed paraffin-embedded tissue sections were deparaffinised, washed and antigen retrieval was performed by microwave heating in citrate buffer. For Cox-2 immunostaining an anti-COX-2 rabbit polyclonal primary antibody was used (Cayman, 160126) at a 1:150 dilution, overnight at 4 °C with an anti-rabbit Alexa Fluor 488 secondary antibody at a 1:1,000 dilution, for 2 h at room temperature. Immunostaining for epithelial lineage markers was performed with conjugated antibodies against lysozyme (FITC-conjugated rabbit polyclonal, DAKO EC 3.2.1.17, 1:100) and Dcl1 (Alexa Fluor 647 rabbit monoclonal, EPR6085, Abcam ab202755, 1:400) and primary antibodies against chromogranin-A (rabbit polyclonal, Abcam ab15160, 1:300) and Olfm4 (rabbit monoclonal, D6Y5A, Cell Signaling 39141, 1:300) followed by an anti-rabbit Alexa Fluor 488 secondary antibody as above. Immunostaining for Yap was performed with a rabbit monoclonal primary antibody (D8H1X, Cell Signaling 14074, 1:50, overnight at 4 °C), a goat anti-rabbit biotinylated IgG secondary antibody (Vector, 1:750) and Streptavidin–Alexa Fluor 488 (1:800). Colocalization of Yap–Alexa Fluor

Article

488 and DAPI was detected with the colocalization mode of Volocity software (PerkinElmer). Immunostaining for laminin A1 was performed with a rat monoclonal antibody (AL-4, R&D MAB4656, 1:50) overnight at 4 °C and an anti-rat Alexa Fluor 594 secondary antibody at a 1:1,500 dilution, for 1 h at room temperature. The numbers of positive cells for each epithelial marker per well-oriented crypt or crypt-villus unit were quantified in a blinded fashion.

BrdU

Administration of BrdU (Sigma) was performed intraperitoneally at a dose of 100 µg per g of body weight 2 h before mice were euthanized. BrdU immunohistochemistry was performed in sections of formalin-fixed paraffin-embedded tissues with the BrdU In-Situ Detection Kit (BD Pharmingen). The sections were counterstained with haematoxylin and analysed with a Leica DMI6000B microscope equipped with the Leica Application Suite LAS v.2.7 software. The number of BrdU⁺ cells per well-oriented crypt was quantified in a blinded fashion.

Alkaline phosphatase

Alkaline phosphatase activity was detected in deparaffinized sections with the Vector Red Alkaline Phosphatase Substrate Kit (Vector, SK-5100) in 200 mM Tris-HCl, pH 8.5 according to the manufacturer's instructions. The sections were counterstained with haematoxylin and mounted with DPX or with Cytoseal Xyl (Thermo).

In situ hybridization

In situ hybridization was performed using the C Multiplex Fluorescent Detection Kit v.2 (ACD Bio) according to the manufacturer's instructions. The colons of eight-week-old wild-type mice were excised, rolled up and immediately frozen in liquid nitrogen before embedding in Tissue-Tek OCT compound (Sakura Finetek). Sections with a thickness of 15 µm were prepared for RNAscope analysis using a mouse *Rspo1* probe (ACD Bio 401991), a mouse *Ppib* positive control (ACD Bio 313911) and the bacterial *DapB* probe as a negative control (ACD Bio 310043). DAPI was used as a nuclear counterstain.

Evaluation of tumorigenesis in *Apc*^{Min/+} mice

Early tumorigenesis in *Apc*^{Min/+} mice was examined at the age of five weeks. The entire small intestine was collected and fixed in 10% neutral-buffered formalin solution as a Swiss roll. H&E-stained paraffin sections were examined with a Nikon Eclipse E800 microscope or a Leica DMI6000 B microscope. The number of microadenomas was quantified in sections stained for β-catenin (Fig. 2) or BrdU (Fig. 4). At the age of 5.5 months, tumorigenesis in *Apc*^{Min/+} mice was evaluated in the small intestine and the colon. The small intestine was partitioned into three parts of equal length (duodenum, jejunum and ileum). The tissues were opened longitudinally and the number of macroscopic tumours was quantified. The opened small intestine was rolled, fixed in formalin and H&E-stained paraffin sections were obtained. Pictures of all adenomas detected per section were obtained and their maximal diameter was measured by using the ImageJ software or the Leica Application Suite LAS v.2.7. All analyses were blinded to mouse genotype.

Azoxymethane-induced colon tumorigenesis

Mice were injected intraperitoneally with 10 mg kg⁻¹ of azoxymethane (Sigma, A5486) once per week for 10 weeks starting at the age of 6 weeks. Mice were euthanized at the age of 28 weeks. Dysplasia development and adenoma formation were evaluated in H&E-stained paraffin sections of the colon. RNA was extracted from formalin-fixed, paraffin-embedded tissues with the RecoverAll Total Nucleic Acid Isolation Kit (Thermo Fisher, AM1975).

Prostanoid analysis by HPLC-MS/MS

Prostanoids were extracted with acetone followed by liquid/liquid extraction as previously described³⁶, with some modifications.

In brief, 10–50 mg of the ileum was homogenized in 500 µl PBS spiked with 100 µM butylated hydroxytoluene on ice. PGE₂-d4 (Cayman, no. 314010) and PGD₂-d4 (Cayman, no. 312010) were used as internal controls in each sample from the beginning of the extraction procedure at a final concentration of 10 ng ml⁻¹. The samples were deproteinized with acetone. After mixing for 4 min and centrifugation at 2,000g for 10 min at 4 °C, the samples were transferred to clean 15-ml glass tubes, mixed with 800 µl hexane for 30 s and centrifuged for 10 min at 2,000g at 4 °C. The lower phase was acidified to pH 3.5 with formic acid and then mixed with chloroform. After mixing for 30 s and centrifugation for 10 min at 2,000g at 4 °C, the lower chloroform phase was evaporated to dryness under a stream of nitrogen and redissolved in 50 µl of methanol. HPLC-MS/MS analysis was performed using a modification of a method previously described³⁷. From each sample a volume of 5 µl was injected into a Gemini 5 µm C18 110 Å, 100 × 2 mm HPLC column (Phenomenex, OOD-4435-BO) coupled with an Agilent 6490 QQQ Triple Quadrupole mass spectrometer with electrospray ionization in negative mode (Yale West Campus Analytical Core). The flow rate was 0.2 ml min⁻¹ and the column was maintained at ambient temperature. The analysis was performed using an acetonitrile-based gradient system mixing two solvents: solvent A was acetonitrile/water/glacial acetic acid, 45/55/0.02 (v/v/v); solvent B was acetonitrile/water/glacial acetic acid, 90/10/0.02 (v/v/v). The analytes were separated using the following gradient: 0.0–8.0 min, 0% solvent B; 8.0–8.1 min, 0 to 50% solvent B; 8.0–12.0 min 50% solvent B; 12.0–12.1 min, 50 to 70% solvent B; 12.1–20.0 min 70% solvent B; 20.0–20.1 min, 70 to 0% solvent B; 20.1–30.0 min 0% solvent B. The capillary voltage was set at 3,500 V, source temperature at 120 °C, desolvation temperature at 360 °C and cone voltage at 35 V. The detection of prostanoids was based on the multiple reaction monitoring (MRM) method. The transition of precursor masses to specific fragments was monitored using a collision energy of 25–30 eV. PGE₂ and PGD₂ which have a similar MRM mass transition (*m/z* 351 → 271) and PGE₂-d4/PGD₂-d4 which also have similar MRM mass transition (*m/z* 355 → 275) were distinguished on the basis of their different elution time from the HPLC column. The MRM mass transition for PGF_{2α} was *m/z* 353 → 193 and for PGI₂ was *m/z* 351 → 215. The data were analysed with the Agilent MassHunter Workstation software, v.B.07.00. For each mass transition the area under the curve was normalized with that of the corresponding internal labelled control and a relative abundance was calculated. The relative abundances calculated were normalized based on the weight of the tissue sample. PGD₂-d4 was used as a control for PGD₂ and its metabolites. PGE₂-d4 was used as a control for the rest of the prostanoids.

Single-cell RNA sequencing and data analysis

Single-cell RNA-seq was performed with the Drop-seq protocol as described previously^{38,39} with minor modifications. Drop-seq analysis of mesenchymal/lamina propria cells isolated from the middle and distal colon of wild type mice was performed in two biological replicates. For each biological replicate, the colons of *n* = 2 mice were pooled. The vast majority of intestinal epithelial cells were depleted by EDTA treatment as described above. *N* = 5 Drop-seq collections were processed in total, two from the first biological replicate and three from the second (Extended Data Fig. 1e). Drop-seq analysis of Ptger4-OFF and Ptger4-ON crypt/fibroblast co-cultures was performed on day 4 of the protocol in one pool of six Ptger4-ON co-cultures and one pool of six Ptger4-OFF co-cultures with one Drop-seq collection per pool. For Drop-seq analysis of crypts, epithelial cells isolated from the small intestine of *Ptger4*^{fl/fl} and *Ptger4*^{ΔIEC} mice, tissues from *n* = 2 mice per genotype were independently processed as biological replicates. A total of three Drop-seq collections were processed for each genotype, two from the first biological replicate and one from the second (Extended Data Fig. 8c).

The cells were diluted to a concentration of 100 cells per µl and 1-ml aliquots were used as input for each collection of the Drop-seq protocol^{38,39}. The beads were purchased from ChemGenes

(no. Macosko201110) and the polydimethylsiloxane co-flow microfluidic droplet generation device was generated by Nanoshift. Samples were processed for cDNA amplification within ~15 min of collection. Populations of 5,000 beads (~150 cells) were separately amplified for 15 cycles of PCR and pairs of PCR products were co-purified by the addition of 0.6× AMPure XP beads (Agencourt). Libraries were prepared and tagged by Nextera XT using 1,000 pg of cDNA input, the custom primer P5_TSO_Hybrid³⁸ and Nextera XT primers N701-N705 (Illumina). Libraries from intestinal mesenchymal cells and crypt epithelial cells were sequenced on the Illumina HiSeq platform (paired end, 2 × 150 bp) and libraries from crypt–fibroblast co-cultures on the Illumina NextSeq 500 platform (paired end, read 1 20 bp; read 2 60 bp), using a Read1CustomSeqB³⁸ primer for read 1.

Single-cell RNA-seq data were processed as described³⁸ to generate a digital expression matrix with transcript count data. This matrix was filtered retaining cells with more than 1,000 transcripts and less than 10% mitochondria transcripts. We then log transformed each entry of the matrix by computing $\log(\text{TPM}/100 + 1)$, where TPM is transcripts per million (meaning that the sum of all gene levels is equal 1,000,000). After normalization, we used adaptively thresholded low rank approximation (ALRA)⁴⁰ to impute the matrix and fill in the technical dropped-out values. Subsequently, to visualize the cell subpopulations in two dimensions, we applied principal component analysis followed by *t*-SNE⁴¹, a nonlinear dimension reduction method, to the log-transformed data. DBSCAN⁴² and graph-based clustering (Seurat, Satija lab) were then used to generate clusters that were overlaid on the *t*-SNE coordinates to investigate cell subpopulations. Marker genes for each cluster of cells were identified using the Wilcoxon test with Seurat. Pathway enrichment analysis was performed by GSVA⁴³ and *P* values were calculated with the moderated *t*-test implemented in the Limma R package. For the adjusted *P* values the false discovery rate (Benjamini–Hochberg) correction method was used.

In the fibroblast–crypt co-culture, single-cell RNA-seq experiment, epithelial cells were distinguished from fibroblasts and selected on the basis of unbiased clustering and known marker genes as shown in the Extended Data Fig. 4d. Cell-cycle analysis was adapted from Seurat. First, a score was calculated for each cell on the basis of the expression of G2M and S phase markers. Then, a discrete classification of cell cycle was assigned to each cell by comparing its G2M and S scores. Cells expressing neither were classified into the G1 group because they are less likely to be cycling.

A metagene score was assigned on the basis of publicly available bulk and single-cell RNA-seq datasets. For each of these datasets we selected significantly differentially expressed genes and constructed a metagene defined as weighted average of the log-transformed expression of these differentially expressed genes with weights equal to the log fold ratio of these genes in the respective dataset. More specifically, if we assume we have a metagene *M* that contains *m* genes: {gene₁, gene₂, ..., gene_m} and each gene *i* has log fold change (FC_{*i*}) in the data we use for the signature of interest, and each gene *i* has an expression value of *x*_{gene_{*i*}} in a given cell in our dataset, then the score for *M* in this specific cell is calculated as:

$$S_M = \sum_{i=1}^m x_{\text{gene}_i} \times \log \text{FC}_i$$

Each cell from our single-cell dataset was characterized by a score associated with each of the metagenes. The extent of differential behaviour between distributions of the total cell populations of two different conditions (Ptger4-ON vs Ptger4-OFF cells) was assessed for each metagene using the Kolmogorov–Smirnov test. We built metagenes for stem cells, enterocytes, Paneth cells, goblet cells, enteroendocrine cells and tuft cells by using lists of population-specific genes based upon plate single-cell RNA-seq data from the mouse intestinal epithelium¹⁵. We also built the following metagenes: (1) a β-catenin program metagene

based on bulk RNA-seq data from organoids bearing a murine stabilized mutant *Ctnnb1*^{stab} transgene and normal organoids (GSE93947), (2) A Yap program metagene based on bulk RNA-seq data from Yap-over-expressing and normal crypts isolated from doxycycline-treated and untreated *YapTg* inducible transgenic mice respectively (GSE66567)¹⁸, (3) An early *Apc*^{Min/+} tumorigenesis program metagene based on microarray gene-expression data from the nonpolypotic sections of terminal ileum from *Apc*^{Min/+} and wild-type mice (GSE49970).

Single-cell RNA-seq data of the healthy human colonic mesenchyme³ were obtained from GSE114374. Single-cell RNA-seq data of regenerating intestinal crypts¹⁶ were obtained from GSE117783. These datasets were processed by ALRA and the subsequent steps as above.

Abdominal irradiation of mice

For abdominal X-ray irradiation an X-RAD 320 Biological Irradiator (Precision X-ray) was used (Research Irradiator Facilities, Department of Therapeutic Radiology, Yale School of Medicine). Mice were anaesthetized and irradiated individually. 15 mm-thick lead was used for head, limb and tail shielding. Mice were treated at a distance of 50 cm from the radiation source with 320 kV, 12.5 mA X-rays, using a filter consisting of 2.0 mm Al. The mouse abdomen was centred between a 55 mm × 65 mm target window outlined by an adjustable collimator. The dose rate was measured with an ionization chamber by members of the Radiation Safety Division at Yale University. The abdominal dose rate was 235 cGy min^{−1}. Average dose to shielded areas was 3.54 cGy min^{−1}.

Flow cytometry and sorting

Freshly isolated stromal cells from *Col6-cre-Rosa26*^{tdTomato/+} mice were stained with monoclonal antibodies against CD45 (Biolegend) and used for FACS sorting. FACS sorting was performed at the Yale Flow Cytometry Facility with a BD FACS Aria II sorter equipped with FACSDiva 7 software. Freshly isolated stromal cells from *Col6-cre-Rosa26*^{mT/mG} *Ptgs2*^{fl/fl} mice were sorted on the basis of their GFP and tdTomato fluorescent protein expression with a BD FACS Aria III sorter (BD) equipped with FACSDiva software at the Flow Cytometry Facility of BSRC Fleming. Single-cell suspensions from organoid cultures and co-cultures were obtained as described above, stained with monoclonal antibodies against Cd24 (Clone M1/69, Biolegend) and Sca-1 (Clone D7, Biolegend) and analysed at the Yale Flow Cytometry Facility with a BD LSR II cytometer equipped with FACSDiva software. Data analysis was performed with the FlowJo software.

Statistical analysis

Statistical analyses were performed with GraphPad Prism 7.01. Normality was tested with the Shapiro–Wilk *W* test. For *W* < 0.05, differences in means were tested for statistical significance with two-tailed Mann–Whitney test or Kruskal–Wallis test. For *W* > 0.05, variances were compared by *F* test and if similar (*F* test, *P* > 0.05), unpaired two-tailed Student's *t*-test or one-way ANOVA was applied, otherwise (*F* test, *P* < 0.05) unpaired two-tailed Welch's *t*-test was applied. For paired comparisons, statistical significance was tested with paired *t*-test for *W* > 0.05 or with Wilcoxon matched-pairs signed-rank test for *W* < 0.05. *P* values < 0.05 were considered as statistically significant. Survival curves were compared by the log-rank test using GraphPad Prism 7.01.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data that support the findings of this study are available within the paper and its Supplementary Information files. All Drop-seq data

Article

that support the findings of this study have been deposited in the Gene Expression Omnibus (GEO) repository with the accession code GSE142431.

Code availability

The code used for single-cell RNA-seq data analysis is available in GitHub (https://github.com/KlugerLab/Scripts_Roulis_et_al_2020).

27. Ishikawa, T. O. & Herschman, H. R. Conditional knockout mouse for tissue-specific disruption of the cyclooxygenase-2 (*Cox-2*) gene. *Genesis* **44**, 143–149 (2006).
28. Armaka, M. et al. Mesenchymal cell targeting by TNF as a common pathogenic principle in chronic inflammatory joint and intestinal diseases. *J. Exp. Med.* **205**, 331–337 (2008).
29. Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45**, 593–605 (2007).
30. Moser, A. R., Pitot, H. C. & Dove, W. F. A dominant mutation that predisposes to multiple intestinal neoplasia in the mouse. *Science* **247**, 322–324 (1990).
31. Schneider, A. et al. Generation of a conditional allele of the mouse prostaglandin EP4 receptor. *Genesis* **40**, 7–14 (2004).
32. Zhang, N. et al. The Merlin/NF2 tumor suppressor functions through the YAP oncoprotein to regulate tissue homeostasis in mammals. *Dev. Cell* **19**, 27–38 (2010).
33. Madison, B. B. et al. *Cis* elements of the villin gene control expression in restricted domains of the vertical (crypt) and horizontal (duodenum, cecum) axes of the intestine. *J. Biol. Chem.* **277**, 33275–33283 (2002).
34. Barker, N. et al. Identification of stem cells in small intestine and colon by marker gene *Lgr5*. *Nature* **449**, 1003–1007 (2007).
35. Madisen, L. et al. A robust and high-throughput Cre reporting and characterization system for the whole mouse brain. *Nat. Neurosci.* **13**, 133–140 (2010).
36. Roulis, M. et al. Intestinal myofibroblast-specific Tpl2–Cox-2–PGE2 pathway links innate sensing to epithelial homeostasis. *Proc. Natl Acad. Sci. USA* **111**, E4658–E4667 (2014).
37. Masoodi, M. & Nicolaou, A. Lipidomic analysis of twenty-seven prostanoids and isoprostanes by liquid chromatography/electrospray tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **20**, 3023–3029 (2006).
38. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
39. Shekhar, K. et al. Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics. *Cell* **166**, 1308–1323 (2016).

40. Linderman, G. C., Zhao, J. & Kluger, Y. Zero-preserving imputation of scRNA-seq data using low-rank approximation. Preprint at <https://www.biorxiv.org/content/10.1101/397588v1> (2018).
41. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
42. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD* **96**, 226–231 (1996).
43. Hänzelmann, S., Castelo, R. & Guinney, J. GSEA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* **14**, 7 (2013).

Acknowledgements We thank C. Lieber, J. Alderman and E. Hughes-Picard for administrative assistance; the Yale Pathology Tissue Service–Tissue Procurement and Distribution Facility for providing human tissue samples; D. Gonzalez for assistance in two-photon imaging; M. Graham for assistance in electron microscopy; M. Samiotaki and T. Wu for assistance in mass spectrometry; and Flavell laboratory members R. Jackson and W. Bailis for discussions. M.R. is supported by a Crohn's and Colitis Foundation Career Development Award (510777); M.S., L.-S.F., M.S.K. and M.B. were supported by an Austrian Marshall Plan Foundation Master's Fellowship. This work was supported in part by ERC project MCs-inTEST (340217) (G.K.), the National Natural Science Foundation of China (31930035, 91942311) (B.S.), the Blavatnik Family Foundation and the Howard Hughes Medical Institute (R.A.F.).

Author contributions M.R. conceived and designed the study and wrote the manuscript. M.R. designed, performed and analysed experiments in Figs. 1–4 and Extended Data Figs. 1–10, assisted by M.S., L.-S.F., M.S.K., M.B., H.N.B. and J.R.B.; A.K. performed experiments in Fig. 2 and Extended Data Figs. 3, 9, assisted by V.K., N.C. and A.H. P.B. implemented Drop-seq. J.Z., R.Q. and Y.K. analysed Drop-seq data. E.K. and V.A. implemented HPLC–MS/MS analyses. X.Z. and B.S. performed in situ hybridization, H.R.H. and J.J. contributed *Ptgs2^{fl/fl}* and *Ptgs2^{SL}* mice, R.M.B. contributed *Ptger4^{fl/fl}* mice and D.J. contributed human FFPE tissues. G.K. and R.A.F. supervised all research, participated in the interpretation of results and edited the manuscript.

Competing interests R.A.F. is a scientific advisor to GlaxoSmithKline and a shareholder and consultant for Zai Lab. All other authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2166-3>.

Correspondence and requests for materials should be addressed to M.R., G.K. or R.A.F.

Peer review information *Nature* thanks Garret A. FitzGerald, Dominic Grün, Kun-Liang Guan, Don W. Powell, Omer Yilmaz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.

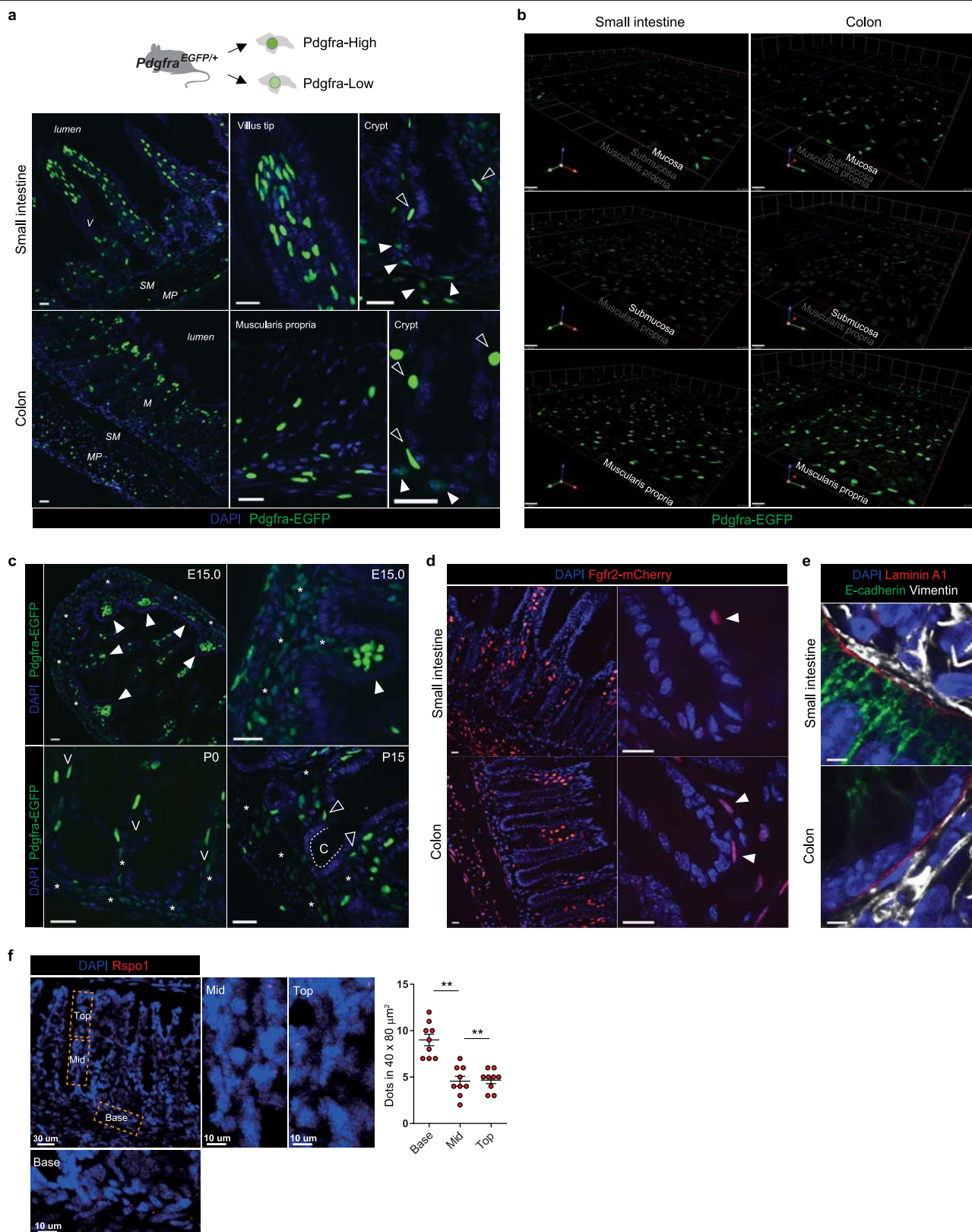


Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | Mesenchymal *Ptgs2* expression in the

microenvironment of crypt stem cells. **a**, Transmission electron microscopy photograph of the base of a mouse ileum crypt. P, Paneth cell; S, columnar basal stem cell; F, fibroblast. Scale bar, 5 μ m. Indicative of independent observations in two experiments. **b**, Immunostaining for Lgr5-eGFP and Vimentin in the ileum of an *Lgr5-eGFP-IRES-creERT2* mouse. Scale bar, 20 μ m. Indicative of independent observations in one experiment. **c**, *PTGS2* relative gene expression (RE) in intestinal epithelial cells (IECs) and stromal cells isolated from healthy human colonic tissues ($n = 6$ individuals). Statistical comparison performed with two-tailed Wilcoxon matched-pairs signed-rank test. **d**, *Ptgs2* gene expression in IECs and stromal cells isolated from the ileum and the colon of wild-type mice ($n = 4$). Statistical significance was determined by two-tailed paired *t*-test. **e**, Biological replicates of the Drop-seq experiment shown in Fig. 1a visualized on the respective *t*-SNE plot depicting $n = 3,179$ mesenchymal cells. Mesenchymal cells were independently isolated from two groups of wild-type mice (biological replicates 1 and 2). From each of these isolations up to three independent Drop-seq samples were collected (A to C) for a total of

five samples. **f**, All *Ptgs2*-expressing single cells ($n = 1,136$) detected in the experiment shown in Fig. 1a, c were analysed separately and re-clustered. Cluster annotations are visualized on a *t*-SNE plot. Violin plots display the entire distribution of gene expression levels per single cell in each cluster for key mesenchymal marker genes. F, fibroblasts. **g**, Schematic representation of the arachidonic acid metabolism pathway. For each mesenchymal cluster shown in Fig. 1a, violin plots display the entire distribution of gene expression levels per single cell for six genes involved in the metabolism of arachidonic acid to prostanoids. Data from $n = 3,179$ single mesenchymal cells are shown. **h**, Analysis of single-cell RNA-seq data (GSE11434) from the healthy human colonic mesenchyme³. Clustering results for $n = 4,348$ cells and cluster annotations are visualized on a *t*-SNE plot. The annotations of stromal populations are matched with the ones reported by Kinchen et al.³ on the basis of the respective markers. Expression levels of *PTGS2* per single cell are visualized on a *t*-SNE plot. The entire range of gene expression levels per single cell for *PTGS1*, *PTGS2* and key mesenchymal marker genes is displayed in violin plots. Data are mean \pm s.e.m.; ns, non-significant; * $P < 0.05$; ** $P < 0.01$.

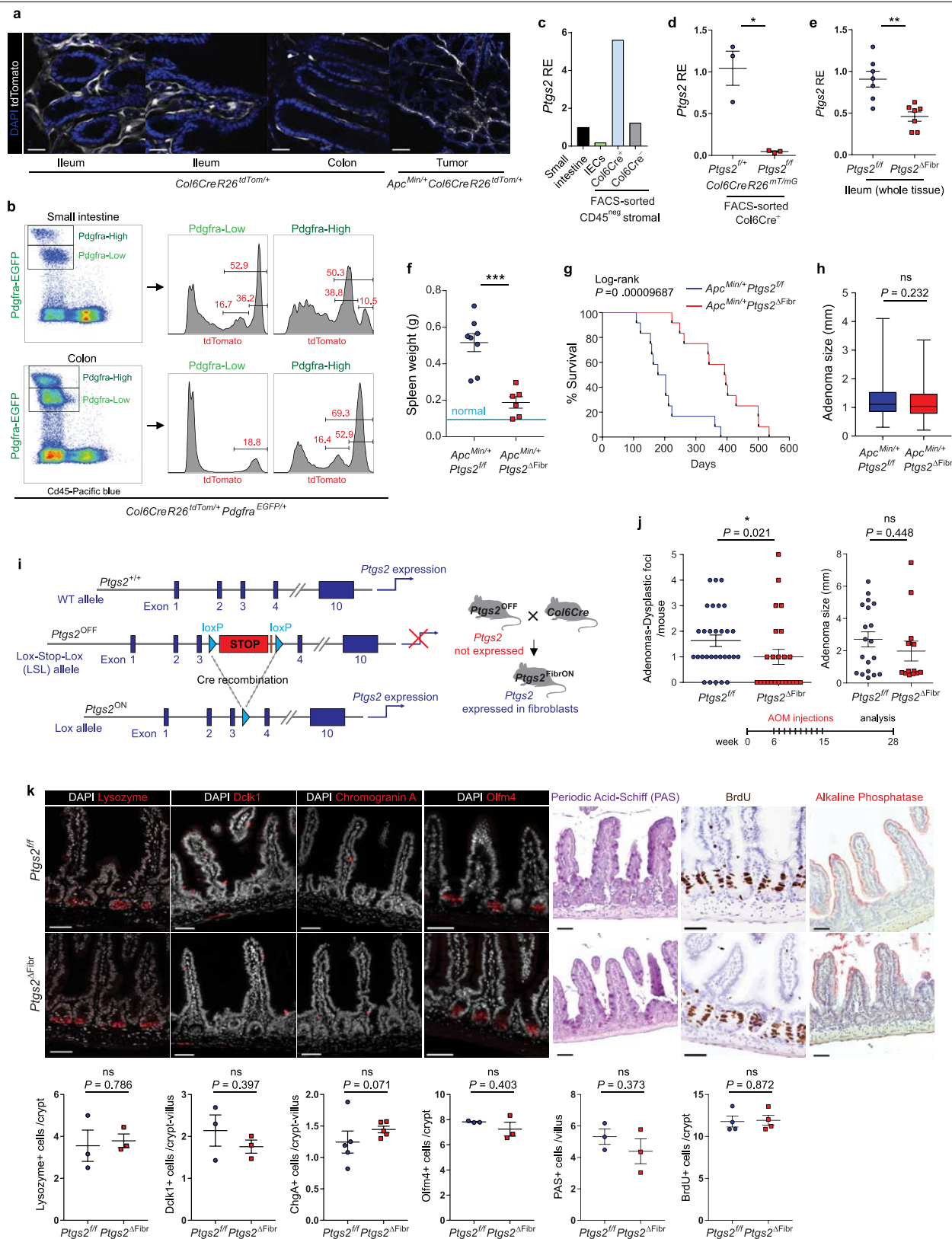


Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | Location of major fibroblast populations in the mouse intestine. **a**, Detection of *Pdgfra*-expressing mesenchymal cells in the intestine of adult *Pdgfra-H2B-eGFP*-knockin mice⁴. Two distinct populations of *Pdgfra*^{high} and *Pdgfra*^{low} mesenchymal cells were detected in fixed tissue sections by direct eGFP fluorescence (green) and confocal microscopy. Nuclei are stained with DAPI (blue). *Pdgfra*^{high} cells are located under the epithelium along the crypt–villus axis and in the muscularis propria. They form clusters at the tips of villi and the apical part of the colonic mucosa. *Pdgfra*^{low} cells are located in the inner part of the villi, the pericryptal area and the submucosa. Filled arrows indicate pericryptal *Pdgfra*^{low} cells. Open arrows indicate subepithelial *Pdgfra*^{high} cells. M, mucosa; V, villus; SM, submucosa; MP, muscularis propria. Scale bars, 20 μ m. Data are representative of six independent experiments.

b, Detection of *Pdgfra*^{high} and *Pdgfra*^{low} fibroblasts in the fresh, intact intestine of adult *Pdgfra-H2B-eGFP*-knockin mice⁴ by two-photon microscopy. The cells were detected by direct eGFP fluorescence (green). *Pdgfra*^{high} cells are predominant in the muscularis propria, whereas *Pdgfra*^{low} cells are predominant in the submucosa. Both populations are present in the mucosa. Data are representative of independent observations from one experiment. Scale bars, 100 μ m. **c**, Detection of *Pdgfra*^{high} and *Pdgfra*^{low} fibroblasts in the intestine of *Pdgfra-H2B-eGFP*-knockin embryos on embryonic day 15 (E15.0)

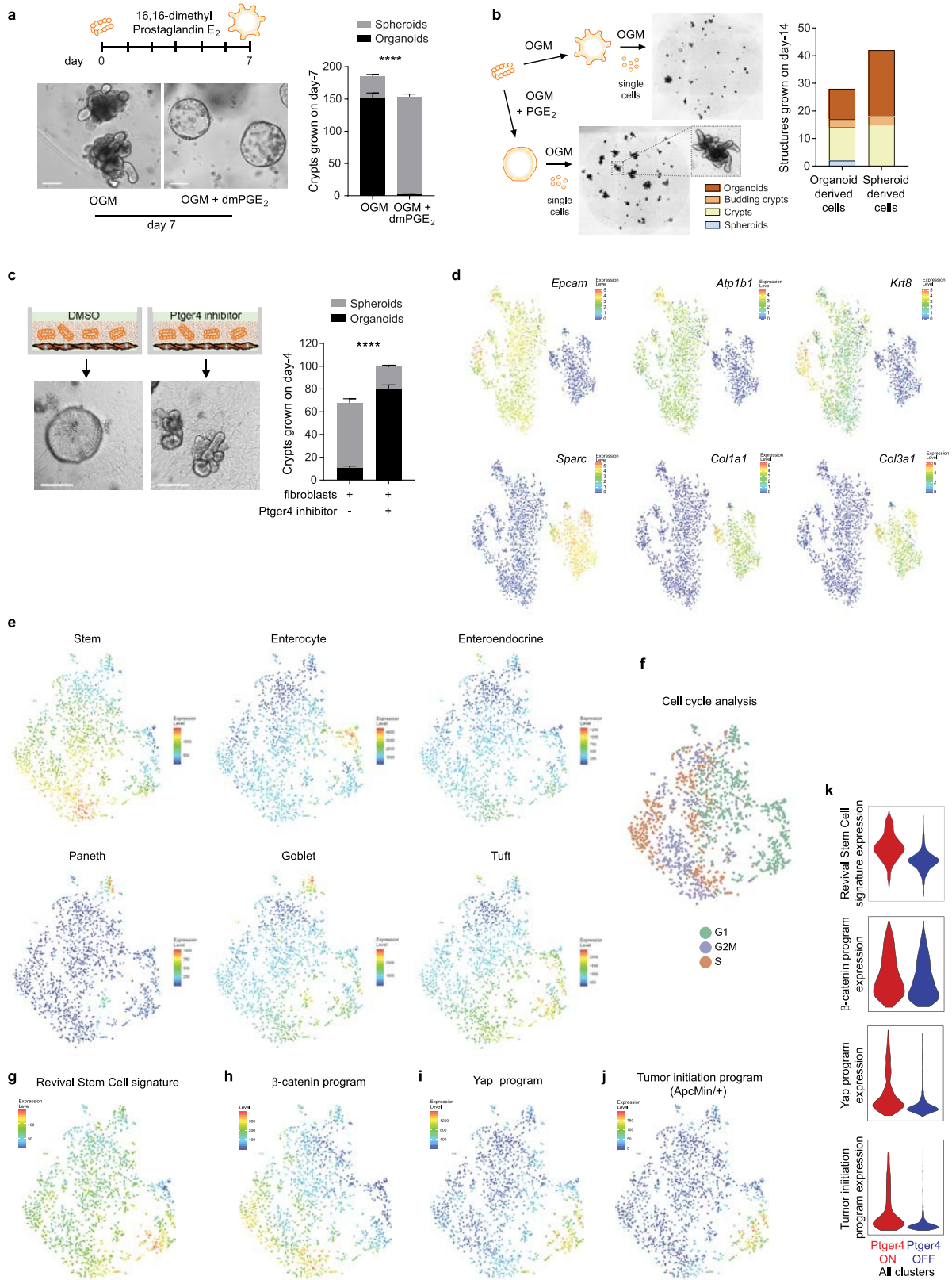
and in early postnatal development. E15.0: clusters of *Pdgfra*^{high} cells in early villi are indicated by white arrows. *Pdgfra*^{low} mesenchymal cells occupy the rest of the mesenchyme (asterisks). P0: *Pdgfra*^{high} cells are observed in the villi (V) and *Pdgfra*^{low} cells are observed both in the villi and in the rest of the mesenchyme (asterisks). P15: *Pdgfra*^{low} cells surround an early crypt (C) and *Pdgfra*^{high} cells are located at the edges of the crypt (open white arrows). *Pdgfra*^{low} cells occupy the inner mesenchyme (asterisks). Data are representative of independent observations from one experiment per developmental stage. Scale bars, 20 μ m. **d**, The location of *Fgfr2*-expressing mesenchymal cells was determined in the intestine of an *Fgfr2-T2A-H2B-mCherry*-knockin mouse⁵, by detecting direct mCherry fluorescence (red) in the nucleus (blue, DAPI). The arrows indicate pericryptal *Fgfr2*⁺ fibroblasts. Data are representative of independent observations from one experiment. Scale bars, 20 μ m. **e**, Immunostaining for laminin A1 (encoded by *Lama1*), the epithelial marker E-cadherin and the mesenchymal marker vimentin in the normal mouse intestines shows that laminin A1 is detected specifically at the mesenchymal–epithelial interface. Data are representative of two independent experiments. Scale bars, 5 μ m. **f**, In-situ hybridization analysis showing the location of *Rspo1*-expressing cells in the normal mouse colon. The position of *Rspo1*-expressing cells along the crypt axis was quantified in $40 \times 80 \mu\text{m}^2$ sub-epithelial areas at the base, middle and top sections of $n = 9$



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Mice with fibroblast-specific ablation or fibroblast-restricted expression of Cox-2. **a**, Immunofluorescence of ileum and colon sections from *Col6-cre-Rosa26^{tdTomato/+}* mice (scale bar, 20 μ m) and of a small intestinal tumour section from an *Apc^{Min/+}-Col6-cre-Rosa26^{tdTomato/+}* mouse (scale bar, 150 μ m). Data are representative of two experiments. **b**, Efficiency of *Col6-cre*-mediated recombination of a lox-stop-lox tdTomato reporter in *Pdgfra^{high}* and *Pdgfra^{low}* *Cd45⁻* cells determined by flow cytometry in intestinal mesenchymal and lamina propria cells isolated from the small intestine and the colon of *Col6-cre-Rosa26^{tdTomato/+}Pdgfra^{eGFP/+}* mice in one experiment. **c**, *Ptgs2* relative gene expression (RE) in whole tissue, isolated IECs, FACS-sorted *Col6Cre⁺* fibroblasts (*Cd45⁻tdTomato⁺*) and *Col6Cre⁻* mesenchymal cells (*Cd45⁻tdTomato⁻*) from the small intestine of *Col6-cre-Rosa26^{tdTomato/+}* mice ($n = 3$, pooled). Representative of two experiments. **d**, Efficiency of *Col6-cre*-mediated *Ptgs2* gene ablation in *Col6-Cre⁺* mesenchymal cells determined by RT-qPCR analysis of *Ptgs2* expression in FACS-sorted *Col6-cre⁺* fibroblasts (*eGFP⁺*) from the small intestine of *Col6-cre-Rosa26^{mT/mG}Ptgs2^{fl/+}* ($n = 3$) and *Col6-cre-Rosa26^{mT/mG}Ptgs2^{fl/fl}* ($n = 3$) mice. Unpaired two-tailed Welch's *t*-test. **e**, Expression of the *Ptgs2* gene in whole tissue ileum of littermate *Ptgs2^{fl/fl}* and *Ptgs2^{ΔFibr}* mice ($n = 7$ each). Two-tailed *t*-test. **f**, Spleen weight of 5.5-month-old *Apc^{Min/+}Ptgs2^{fl/fl}* ($n = 8$) and *Apc^{Min/+}Ptgs2^{ΔFibr}* ($n = 6$) mice. Average spleen weight of ($n = 6$) normal littermates (*Ptgs2^{fl/fl}*) is displayed for comparison. Two-tailed *t*-test. **g**, Survival analysis of *Apc^{Min/+}Ptgs2^{fl/fl}* ($n = 12$) and *Apc^{Min/+}Ptgs2^{ΔFibr}* ($n = 12$) mice. A two-tailed $P = 0.00009687$ was calculated by log-rank test. **h**, Size of 274 adenomas from 5.5-month-old *Apc^{Min/+}Ptgs2^{fl/fl}* ($n = 16$) and *Apc^{Min/+}Ptgs2^{ΔFibr}* ($n = 18$) mice. The whiskers extend from minimum to maximum and the box

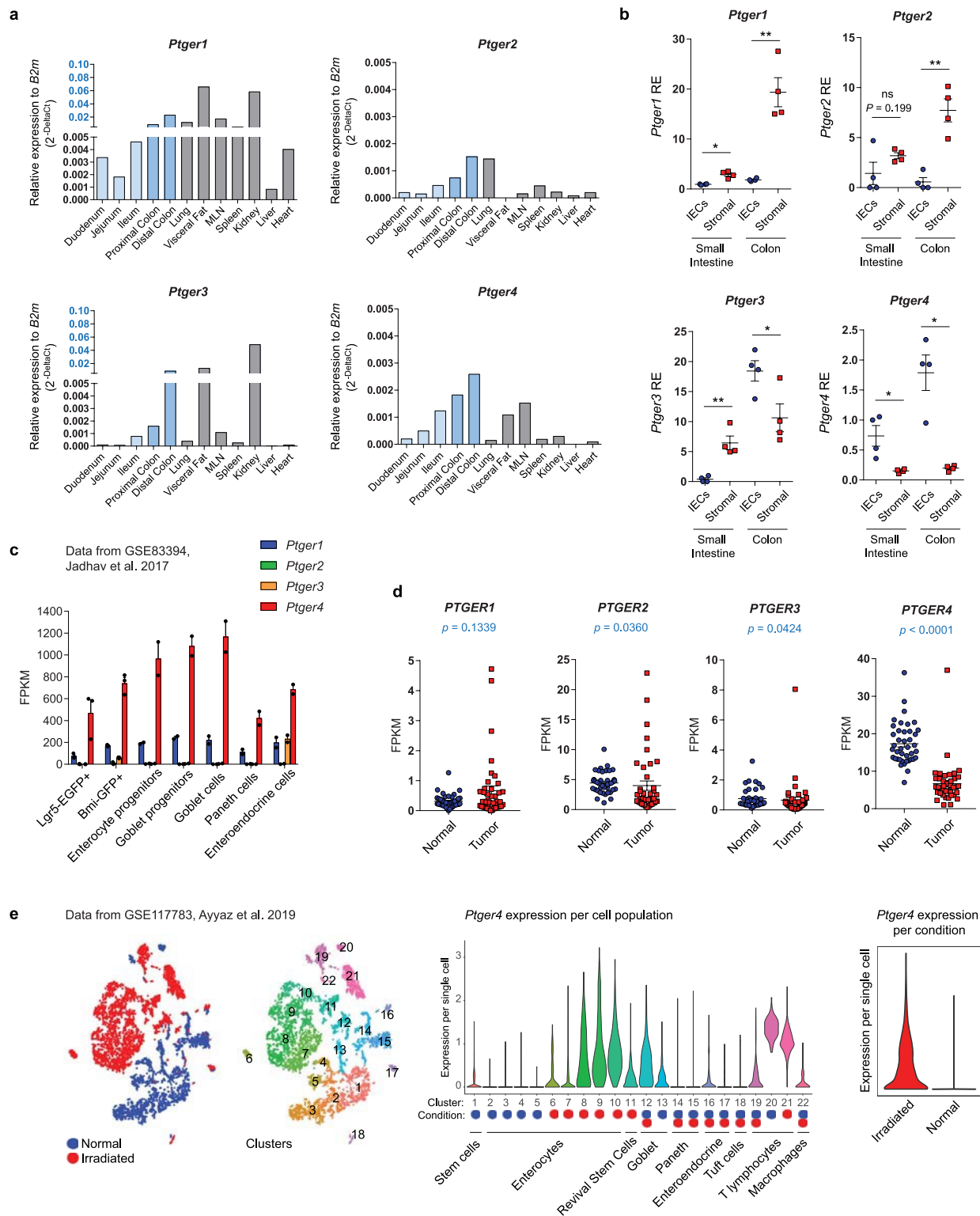
extends from the 25th to 75th percentiles with the median indicated. Two-tailed Mann-Whitney test. **i**, Generation of knockin mice bearing a lox-stop-lox cassette insertion in intron-3 of the *Ptgs2* gene which prevents its expression (*Ptgs2^{OFF}*). *Col6-cre*-mediated excision of the lox-stop-lox cassette reactivates *Ptgs2* expression specifically in fibroblasts (*Ptgs2^{FibrON}*). The orange box depicts an *frt* site remaining from the *flp*-mediated removal of an *frt*-flanked PGK-neomycin selection cassette (see Methods). **j**, *Ptgs2^{fl/fl}* ($n = 30$) and *Ptgs2^{ΔFibr}* ($n = 24$) mice were subjected to 10 weekly intraperitoneal injections with 10 mg kg⁻¹ azoxymethane as displayed. Quantification of the number of dysplastic foci and microadenomas per mouse and quantification of tumour size is shown. Statistical significance was tested by two-tailed Mann-Whitney test. **k**, Quantification of intestinal epithelial populations in the ileum of littermate *Ptgs2^{fl/fl}* and *Ptgs2^{ΔFibr}* mice ($n = 3-5$ per genotype). Immunostaining was performed for markers of Paneth cells (lysozyme), tuft cells (*Dclk1*), enteroendocrine cells (chromogranin A) and stem cells (*Olfm4*). Goblet cells were identified by periodic acid Schiff (PAS) staining and enterocytes were identified by detecting alkaline phosphatase enzymatic activity. Incorporation and immunohistochemical detection of BrdU was used to determine the numbers of cycling cells. Data for each mouse represent mean number of positive cells per crypt or crypt-villus unit as indicated. $N = 400-822$ crypts and/or villi were evaluated per staining. Statistical comparisons were performed with two-tailed unpaired *t*-test except for *Olfm4⁺* cells for which unpaired *t*-test with Welch's correction was applied. Scale bars, 50 μ m. All data represent mean \pm s.e.m. unless otherwise indicated. ns, non-significant; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | PGE₂-driven spheroids contain more functional stem cells. **a**, Crypts isolated from the small intestine of wild-type mice were grown into organoids by 3D culture with OGM or OGM that was supplemented daily with 0.1 μ M 16,16-dimethyl PGE₂ (dmPGE₂). Indicative images and quantification of the absolute numbers of organoids and spheroids grown per 3D structure are shown. $n = 6$ 3D cultures were evaluated per condition. Scale bar, 100 μ m. **b**, Assessment of stem-cell activity in organoids or PGE₂-driven spheroids grown as in **a** by dissociation into single cells and 3D culture in OGM. Growth of crypts and organoids from the same initial number of cells was quantified on day 14. The results are indicative of five independent experiments starting from independent crypt isolations. **c**, Normal crypts were grown into organoids with OGM in a 3D co-culture with primary mouse intestinal fibroblasts with or without 10 μ M ONO-AE3-208 (Ptger4/EP4 inhibitor). Indicative images and quantification of the absolute numbers of organoids and spheroids grown per 3D structure are shown. $n = 6$ 3D co-cultures were evaluated per condition. Scale bar, 200 μ m. **d**, Separation of $n = 2,192$ fibroblasts and epithelial cells in single-cell RNA-seq data from fibroblast-crypt organotypic cultures on the basis of the expression of key

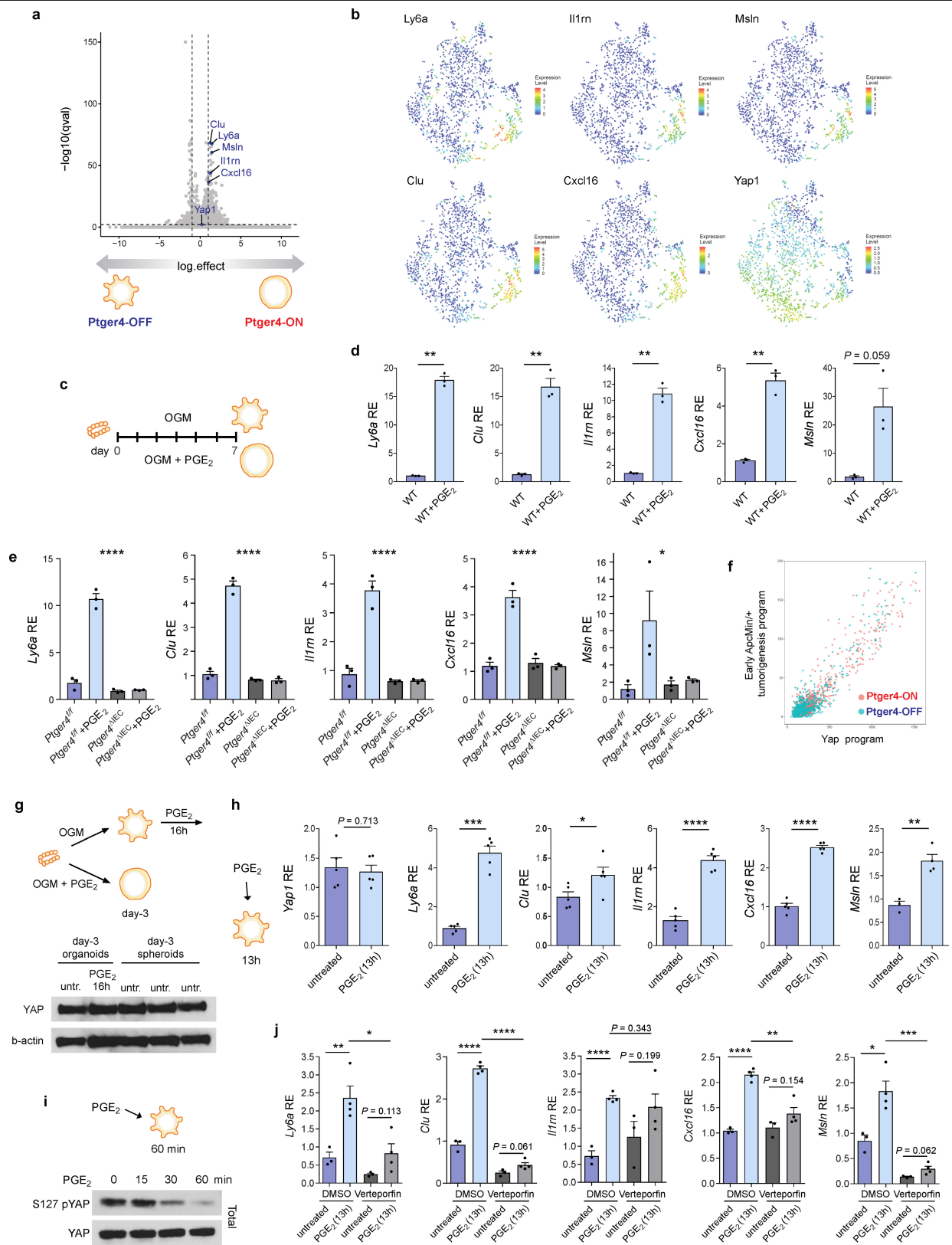
marker genes. Expression of intestinal epithelial marker genes (*Epcam*, *Atp1b1* and *Krt8*) and fibroblast marker genes (*Sparg*, *Col1a1* and *Col3a1*) in single cells from Ptger4-ON and Ptger4-OFF fibroblast-crypt co-cultures is shown projected onto *t*-SNE plots. **e–j**, Single-cell data from Ptger4-ON and Ptger4-OFF fibroblast-crypt co-cultures as shown in Fig. 3d, visualized on the respective *t*-SNE plot depicting $n = 1,585$ epithelial cells. **e**, Expression of epithelial population-specific signatures (metagenes) per single epithelial cell. Population signatures were calculated on the basis of single-cell profiling of the mouse intestinal epithelium¹⁵. **f**, Cell cycle analysis of single epithelial cells projected onto the *t*-SNE plot. **g–j**, Expression levels of metagenes for the signatures or transcriptional programs of RSC¹⁶ (**g**), β -catenin (**h**), Yap (**i**) and early (non-tumour) *Apc*^{Min/+} tumorigenesis (**j**) per single epithelial cell projected onto *t*-SNE plots. **k**, Data from $n = 1,585$ single epithelial cells, visualized in violin plots for each co-culture condition (Ptger4-ON or Ptger4-OFF). The entire range of metagene expression levels per single epithelial cell for the signatures or transcriptional programs of RSCs, β -catenin, Yap and early *Apc*^{Min/+} tumorigenesis is displayed. In **a**, **c**, two-way ANOVA. Data are mean \pm s.e.m. **** $P < 0.0001$.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Expression of PGE₂ receptors in mouse and human tissues. **a**, RT-qPCR analysis for *Ptger1*, *Ptger2*, *Ptger3* and *Ptger4* genes across 12 mouse tissues. Expression relative to *B2m* is displayed as $2^{-\Delta\Delta C_t}$. Data represent one experiment. MLN, mesenteric lymph nodes. **b**, RT-qPCR analysis for *Ptger1*, *Ptger2*, *Ptger3* and *Ptger4* genes in isolated IECs and matched stromal fractions from the small intestine (ileum) and the colon of wild-type mice ($n = 4$). Statistical comparisons were performed by two-tailed paired *t*-test. **c**, Expression levels of *Ptger1*, *Ptger2*, *Ptger3* and *Ptger4* genes determined by RNA-seq in FACS-sorted intestinal epithelial cell populations in 2 or 3 biological replicates and displayed as FPKM (fragments per kilobase of transcript per million mapped reads). Data retrieved from the GSE83394 GEO dataset. **d**, Expression levels of the human *PTGER1*, *PTGER2*, *PTGER3* and *PTGER4* genes in matched normal colon and tumour tissues from colorectal cancer patients

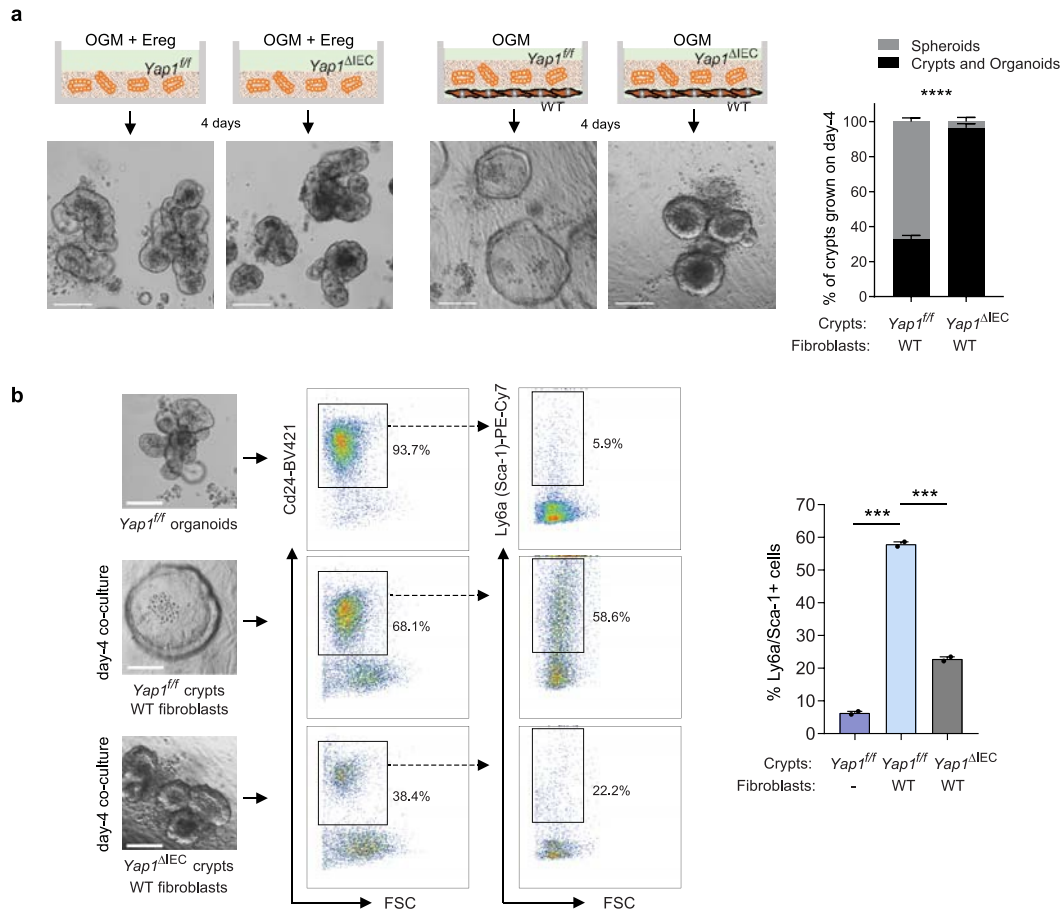
($n = 41$), determined by RNA-seq and displayed as FPKM. Data retrieved from The Cancer Genome Atlas for colon adenocarcinoma (TCGA-COAD dataset). Statistical comparisons were performed by two-tailed Wilcoxon matched-pairs signed-rank test. **e**, Analysis of single-cell RNA-seq data¹⁶ (GSE117783) from crypts isolated from the small intestine of normal mice (blue) and mice treated with 12 Gy irradiation (red). $n = 6,644$ single cells are visualized on *t*-SNE plots based on the experimental condition (normal, $n = 2,882$; irradiated, $n = 3,762$) and the clustering results. Violin plots represent the entire distribution of *Ptger4* expression levels per single cell in each cluster and in each condition. The annotations of epithelial populations are matched with the ones reported by Ayyaz et al.¹⁶ on the basis of the respective markers. **b–d**, Mean \pm s.e.m.; ns, non-significant; * $P < 0.05$; ** $P < 0.01$.



Extended Data Fig. 6 | See next page for caption.

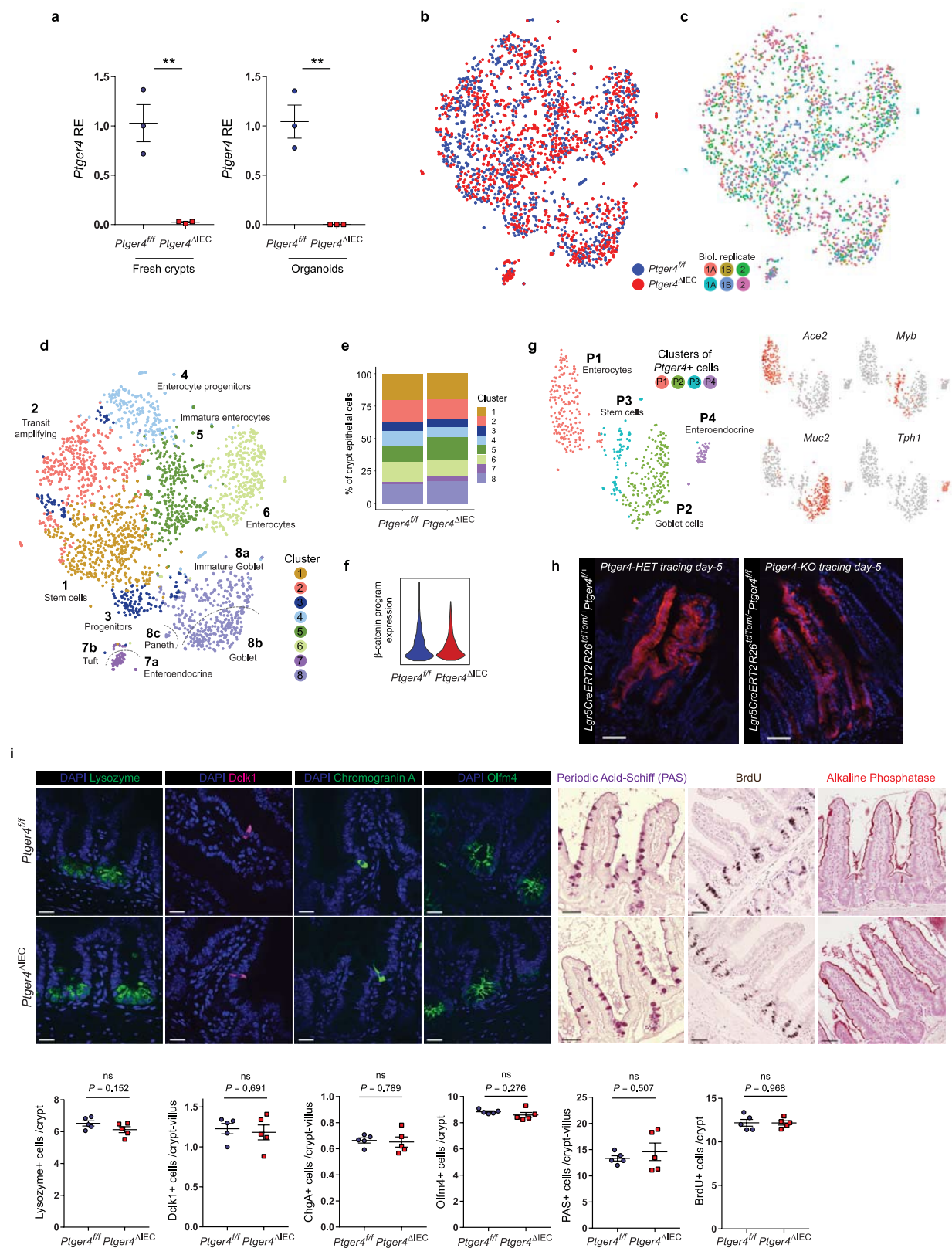
Extended Data Fig. 6 | PGE₂–Ptger4 drive the induction of Yap target genes in intestinal organoids. **a**, Volcano plot displaying the results of differential gene-expression analysis performed in single epithelial cells from Ptger4-ON and Ptger4-OFF fibroblast–crypt co-cultures ($n=1,585$). *Yap1* and Yap target genes¹⁸ are indicated. Moderated t -test with false-discovery rate (Benjamini–Hochberg) correction. **b**, Expression levels of the genes indicated in single epithelial cells from Ptger4-ON and Ptger4-OFF fibroblast–crypt co-cultures ($n=1,585$), projected onto t -SNE plots. **c**, Experimental setup for data shown in **d**, **e**. Crypts were grown into organoids or spheroids by 3D culture in OGM or OGM that was supplemented daily with 0.1 μ M dmPGE₂ for 7 days. Gene expression levels were measured by RT–qPCR on day 7. **d**, Relative expression of Yap target genes (*Ly6a*, *Clu*, *Il1rn*, *Msln* and *Cxcl16*) in day 7 organoids and PGE₂-driven spheroids developed from wild-type crypts. $N=3$ 3D cultures per condition. Two-tailed Welch's t -test. **e**, Relative expression of Yap target genes in day 7 organoids and PGE₂-driven spheroids developed from crypts isolated from *Ptger4*^{fl/fl} and *Ptger4*^{ΔIEC} mice. $n=3$ 3D cultures per genotype and condition. One-way ANOVA. **f**, Correlation between the expression levels of metagenes of a Yap transcriptional program and an early (non-tumour) *Apc*^{Min/+} tumorigenesis transcriptional program in single epithelial cells ($n=1,585$) from

the Ptger4-ON and Ptger4-OFF fibroblast–crypt co-cultures of Fig. 3. **g**, Small intestinal crypts were grown into organoids or spheroids with OGM or OGM that was supplemented daily with 0.1 μ M dmPGE₂. Western blot analysis for Yap1 and β -actin was performed in total lysates from untreated organoids, organoids treated with 0.1 μ M dmPGE₂ for 16 h and untreated spheroids. Data from one organoid and three independent spheroid cultures. **h**, Relative expression of the *Yap1* gene and Yap target genes in wild-type organoid cultures treated with 0.1 μ M dmPGE₂ for 13 h, as determined by RT–qPCR. $n=3$ –5 cultures per condition. Statistical comparisons were performed with unpaired two-tailed t -test. For *Ly6a*, Welch's correction was applied. **i**, Western blot analysis for Ser127 pYap and total Yap performed in total lysates from wild-type organoids stimulated with 0.1 μ M dmPGE₂ for the indicated time-points. Indicative of five independent experiments. **j**, Relative expression of Yap target genes in wild-type organoids treated with 1 μ M verteporfin and 0.1 μ M dmPGE₂ for 13 h. $n=3$ –4 cultures per condition. Statistical comparisons were performed with unpaired two-tailed t -test, two-tailed Welch's t -test or Mann–Whitney test on the basis of the criteria described in Methods. All data are mean \pm s.e.m. * $P<0.05$, ** $P<0.01$, *** $P<0.001$, **** $P<0.0001$.



Extended Data Fig. 7 | Genetic ablation of Yap prevents spheroid formation and Sca-1⁺ stem-cell expansion in fibroblast–crypt organotypic co-cultures. **a**, Crypts isolated from the small intestines of *Yap1^{fl/fl}* and *Yap1^{ΔIEC}* mice were grown into organoids by 3D culture with OGM supplemented with 0.5 mg ml⁻¹ recombinant mouse epiregulin (Ereg) as previously described¹⁸, or in a co-culture with wild-type primary mouse intestinal fibroblasts with OGM without Ereg supplementation. Indicative images and quantification of the percentages of crypts, organoids and spheroids grown per 3D structure are shown. *n* = 2 cultures per condition. Data are representative of two

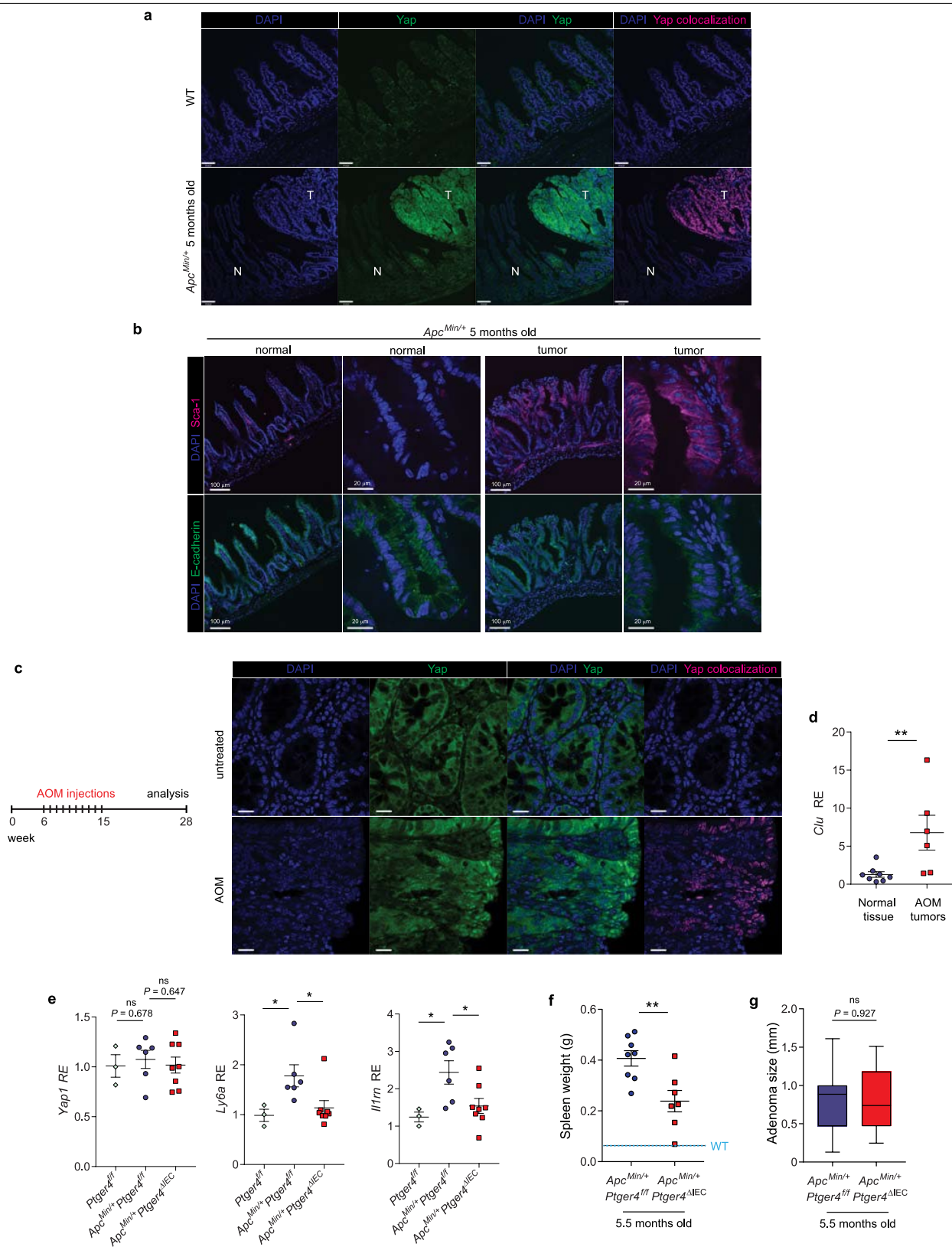
independent experiments. Scale bars, 100 μm. Two-way ANOVA. Data represent mean ± s.e.m. *****P* < 0.0001. **b**, Intestinal crypts isolated from the small intestines of *Yap1^{fl/fl}* and *Yap1^{ΔIEC}* mice were co-cultured with wild-type primary mouse intestinal fibroblasts. On day 4, these co-cultures and control *Yap1^{fl/fl}* organoid cultures were processed into single-cell suspensions and analysed by flow cytometry for Sca-1 expression in Cd24⁺ epithelial cells. *n* = 2 cultures per condition. Scale bars, 100 μm. FSC, forward scatter. Unpaired two-tailed *t*-test. Mean ± s.e.m. ****P* < 0.001.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | *Ptger4* ablation does not affect epithelial lineage differentiation and stem-cell function. **a**, *Ptger4* gene expression in crypts isolated from the ileum of littermate *Ptger4^{fl/fl}* and *Ptger4^{ΔIEC}* mice ($n = 3$ mice per genotype) and in organoids grown from these crypts ($n = 3$ cultures per genotype) determined by RT-qPCR analysis. Two-tailed unpaired *t*-test. **b**, Single-cell RNA-seq (Drop-seq) was performed in crypt epithelial cells isolated from littermate *Ptger4^{fl/fl}* and *Ptger4^{ΔIEC}* mice. Data for 2,439 single epithelial cells are shown in a *t*-SNE plot. **c**, Biological replicates visualized on a *t*-SNE plot. Crypt epithelial cells were independently isolated from two groups of mice per genotype (biological replicates 1 and 2). From the first biological replicate, two independent Drop-seq samples were collected (A and B) for a total number of three samples per genotype. **d**, Clustering and cluster assignments of 2,439 single epithelial cells displayed on a *t*-SNE plot. **e**, Proportion of each epithelial cluster among total crypt epithelial cells in *Ptger4^{fl/fl}* and *Ptger4^{ΔIEC}* mice. **f**, Violin plots showing the entire range of expression levels for a metagene of the β -catenin transcriptional program in $n = 2,439$ single epithelial cells from *Ptger4^{fl/fl}* and *Ptger4^{ΔIEC}* mice. **g**, Analysis of all *Ptger4*-expressing single cells detected ($n = 478$). Re-clustering results of *Ptger4*-expressing single cells with cluster annotations are visualized on a *t*-SNE plot. The expression levels of key marker genes for these clusters are visualized

on *t*-SNE plots. **h**, Lineage tracing of *Ptger4* heterozygous (*Ptger4*-HET) and *Ptger4*-knockout (*Ptger4*-KO) *Lgr5⁺* stem cells. The small intestines of *Lgr5-creERT2-Rosa26^{tdTomato/+}Ptger4^{fl/+}* (*Ptger4*-HET) and *Lgr5-creERT2-Rosa26^{tdTomato/+}Ptger4^{fl/fl}* (*Ptger4*-KO) mice were examined for direct tdTomato fluorescence 5 days after a single injection of 2 mg tamoxifen per mouse. The results shown are representative of independent observations from one experiment. Scale bars, 70 μ m. **i**, Quantification of intestinal epithelial populations in the ileum of littermate *Ptger4^{fl/fl}* ($n = 5$) and *Ptger4^{ΔIEC}* ($n = 5$) mice. Immunostaining was performed for markers of Paneth cells (lysozyme), tuft cells (Dclk1), enteroendocrine cells (chromogranin A) and stem cells (Olfm4). Scale bars, 20 μ m. Goblet cells were identified by PAS staining and enterocytes were detected by alkaline phosphatase enzymatic activity. Scale bars, 50 μ m. Incorporation and immunohistochemical detection of BrdU was used to determine the numbers of cycling cells. Scale bars, 50 μ m. Data for each mouse represent mean number of positive cells per crypt or crypt-villus unit as indicated. $n = 217$ –565 crypts and/or villi were evaluated per staining. Statistical comparisons were performed with two-tailed unpaired *t*-test except for PAS⁺ cells, for which unpaired Welch's *t*-test was applied. Mean \pm s.e.m.; ns, non-significant; ** $P < 0.01$.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Nuclear localization of Yap and activation of Yap target genes in *Apc^{Min/+}* and azoxymethane-induced tumorigenesis.

a, Immunostaining for Yap in the small intestine of five-month-old *Apc^{Min/+}* and wild-type littermate control mice. Nuclear localization of Yap is displayed on the basis of colocalization with DAPI. Normal (N) and tumour (T) areas of the *Apc^{Min/+}* intestine are indicated. Scale bars, 70 μm . Data are indicative of at least ten different tumour areas. **b**, Immunostaining for Sca-1 and the epithelial marker E-cadherin in normal and tumour areas of the small intestine of five-month-old *Apc^{Min/+}* mice. Indicative of two independent experiments. **c**, Immunostaining for Yap in the colon of wild-type mice subjected to 10 weekly intraperitoneal injections with 10 mg kg⁻¹ azoxymethane as indicated and in untreated controls. Nuclear localization of Yap is displayed on the basis of colocalization with DAPI. Scale bars, 20 μm . Data indicative of three mice analysed. **d**, Relative expression of the Yap target gene *Clu* in normal and

tumour areas of the colon of wild-type mice ($n = 8$) subjected to 10 weekly intraperitoneal injections with 10 mg kg⁻¹ azoxymethane as shown in **c**. Two-tailed Mann-Whitney test. **e**, Relative expression of *Yap1* and Yap target genes in the small intestine of 5-week-old *Ptger4^{fl/fl}* ($n = 3$), *Apc^{Min/+} Ptger4^{fl/fl}* ($n = 6$) and *Apc^{Min/+} Ptger4^{ΔIEC}* ($n = 8$) mice. Statistical comparisons were performed with two-tailed *t*-test for *Yap1* and *Il1rn* and with two-tailed Mann-Whitney test for *Ly6a*. **f**, Spleen weight of 5.5-month-old *Apc^{Min/+} Ptger4^{fl/fl}* ($n = 8$) and *Apc^{Min/+} Ptger4^{ΔIEC}* ($n = 7$) mice. Average spleen weight of ($n = 2$) normal littermates (*Ptger4^{fl/fl}*) is displayed for comparison. Two-tailed *t*-test. **g**, Size of 72 adenomas from 5.5-month-old *Apc^{Min/+} Ptger4^{fl/fl}* ($n = 6$) and *Apc^{Min/+} Ptger4^{ΔIEC}* ($n = 4$) mice. The whiskers extend from minimum to maximum and the box extends from the 25th to 75th percentiles with the median indicated. Two-tailed Mann-Whitney test. Mean \pm s.e.m.; ns, non-significant; * $P < 0.05$; ** $P < 0.01$.

localization of Yap is displayed on the basis of colocalization with DAPI. Clearly defined normal (N) and tumour (T) areas are indicated wherever applicable. Images shown are representative of specimens obtained and analysed from $n=16$ patients with the types of colorectal tumours indicated. Patient characteristics and the type of colorectal tumour per individual are described in the Supplementary Table 3. **c**, Schematic representation of the mechanism proposed in the present study. TISC, tumour-initiating stem cell.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|--------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Imaging data were acquired with the Velocity software (PerkinElmer) and the Leica Application Suite LAS V2.7 software. Mass Spectrometry data were collected with the Agilent MassHunter Workstation software, version B.07.00. Real-time PCR data were acquired and with the CFX Manager software (Bio-Rad). Flow cytometry data were acquired with the FACSDiva 7 software.

Data analysis

Imaging data were analyzed with the Velocity software (PerkinElmer) and the Leica Application Suite LAS V2.7 software. Adenoma size was measured in pictures obtained from H/E sections with the ImageJ software or the Leica Application Suite LAS V2.7 Mass Spectrometry data were analyzed with the Agilent MassHunter Workstation software, version B.07.00. Real-time PCR data were analyzed with the CFX Manager software (Bio-Rad). Relative gene expression was calculated with the RelQuant software (Bio-Rad Laboratories). Flow cytometry data were analyzed with the FlowJo V10 software. For statistical analyses and graphs we used GraphPad Prism 7.01 and R. scRNAseq data was analysed with Seurat, Bowtie-2, Picard, GSVA, the Limma R package and custom code which will be available upon request.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data that support the findings of this study are available from the authors upon request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to predetermine sample size for experimentation. For in vivo experiments sample size was determined based upon the availability of mice and was the largest possible and consistent with the known literature. For in vitro experiments with organoids n = 3-5 independent cultures were used per experiment which was sufficient because of the minimal variation observed and consistent with the known literature
Data exclusions	No data were excluded from our analyses
Replication	All experiments were independently reproduced as stated in the figure legends and all attempts of replication were successful. For organoid experiments in mice and humans independent experiments refer to fully independent cultures starting from different mice or different patients respectively
Randomization	There was no need for randomization in our experiments. Mice were analyzed based upon their genotype and no mice were excluded from the analyses.
Blinding	Macroscopic tumor count and all histological analyses (counting of microadenomas, assessment of phenotype after irradiation) were performed in a blinded manner.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input type="checkbox"/>	<input type="checkbox"/> Clinical data

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input type="checkbox"/>	<input checked="" type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

Phospho-YAP (Ser127), clone D9W21, Cell Signaling 13008, Ref 5/18 - lot: 5 and Ref 10/17 - lot: 5, dilution 1:1000 (Western blot); YAP, clone D8H1X, Cell Signaling 14074, dilution 1:50 (IHC-P), 1:1000 (Western blot); TBP, clone D5C9H, Cell Signaling 44059, Ref 4/18 - lot: 1, dilution 1:1000 (Western blot); beta-actin, clone C4, Santa Cruz sc-47778, lot D0615, dilution 1:2000 (Western blot); Akt (pan), clone C67E7, Cell Signaling 4691, Ref 1/19 - lot: 20, dilution 1:1000 (Western blot); Vimentin-Alexa Fluor 647 rabbit monoclonal, clone D21H3, Cell Signaling 9856, Ref 8/19 - lot 13, dilution 1:800 (IHC-P), 1:200 (IF); COX2 rabbit polyclonal, Cayman 160126, lot 0482857-1, dilution 1:150 (IHC-P); GFP-Alexa Fluor 488 rabbit polyclonal, Thermo Fisher A-21311, dilution 1:200 (IF); CD45.2- Pacific Blue™ rat monoclonal, clone 104, Biolegend 109820, lot: B249623, dilution 1:200 (flow); Lysozyme-FITC, rabbit polyclonal, DAKO EC 3.2.1.17, dilution 1:100 (IHC-P); DclK1-Alexa Fluor 647 rabbit monoclonal, clone EPR6085, Abcam ab202755, lot GR229365-2, dilution 1:400 (IHC-P); Chromogranin A rabbit polyclonal, Abcam ab15160, lot ZZG021907A, dilution 1:300 (IHC-P); Olfm4 rabbit monoclonal, clone D6Y5A, Cell Signaling 39141, Ref 12/18 - lot: 1, dilution 1:300 (IHC-P); Ly6a/Sca-1-Alexa Fluor 647 rat monoclonal, clone E13-161.7, Biolegend 122517, lot: B249605, dilution 1:400 (IF); Ly6a/Sca-1-PE-Cy7 rat monoclonal, clone D7, Biolegend 108114, lot: B154904, dilution 1:100 (flow); Cd24 Brilliant Violet 421™ anti-mouse monoclonal, clone M1/69, Biolegend 101825, dilution 1:200 (flow); Laminin A1 rat monoclonal, clone AL-4, R&D MAB4656, dilution 1:50 (IHC-P);

E-cadherin-FITC mouse monoclonal, clone 36/E-Cadherin, BD 612130, lot 8152975, dilution 1:200 (IHC-P)

Validation

All antibodies have been validated by the manufacturer for the species and the application to be used for as described in the data sheets provided.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals

We used the *Mus musculus* C57BL6 strain (laboratory mouse). The generation of *Ptgs2* lox-stop-lox knockin mice is described in this publication. *Ptgs2* lox-stop-lox mice were generated and provided by Harvey R. Herschman, *Col6Cre* mice were provided by George Kollias, *Ptgs2f/f* mice were provided by Harvey R. Herschman, *Ptger4f/f* mice were provided by Richard M. Breyer, *Fgfr2mCherry* mice were provided by Philippe Soriano, *ApcMin/+* mice, *VillinCre* mice, *Lgr5-EGFP-IRES-creERT2* mice, *Rosa26tdTomato/+* mice (Ai14) and *PdgfraEGFP/+* were purchased from the Jackson Laboratories. All mice compared in experiments were littermates, co-housed and sex matched. Both male and female mice were used. Experiments were performed with 8-12 week-old mice unless otherwise stated in the figure legends (tumor experiments at 5 weeks and 5.5 months).

Wild animals

The study did not involve wild animals

Field-collected samples

The study did not involve samples collected from the field

Ethics oversight

All animal experimentation at Yale was performed in compliance with Yale Institutional Animal Care and Use Committee protocols. Experiments in BSRC "Alexander Fleming" were approved by the Institutional Committee of Protocol Evaluation in conjunction with the Veterinary Service Management of the Hellenic Republic Prefecture of Attika according to all current European and national legislation.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics

Human subjects were both men (60%) and women (40%), aged between 23 and 87 years old (average 51.5 years old). Detailed human subject characteristics are provided in the Supplementary Table 3.

Recruitment

All human intestinal tissue samples (fresh or formalin-fixed paraffin-embedded) were derived from patients who underwent surgery at the Yale-New Haven Hospital for the conditions described in the Supplementary Table 3. Tissues were obtained anonymously from the surgical pathology services of the hospital (Yale Pathology Archives) based upon availability without bias in terms of sex and age

Ethics oversight

Yale Human Investigation Committee protocols #0304025173

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about [clinical studies](#)All manuscripts should comply with the ICMJE [guidelines for publication of clinical research](#) and a completed [CONSORT checklist](#) must be included with all submissions.

Clinical trial registration

Provide the trial registration number from [ClinicalTrials.gov](#) or an equivalent agency.

Study protocol

Note where the full trial protocol can be accessed OR if not available, explain why.

Data collection

Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.

Outcomes

Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.

Flow Cytometry

Plots

Confirm that:

- ☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☒ All plots are contour plots with outliers or pseudocolor plots.
- ☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation

For intestinal lamina propria/mesenchymal cell isolation the intestine was dissected, flushed, opened longitudinally and then cut into 1 cm pieces. The tissues were incubated in HBSS containing 1 mM EDTA, 1 mM DTT, 0.2 % FBS, 4-5 times, 10 min each, at 37 °C, 200 rpm. Epithelial cells were depleted by vigorous shaking. After epithelial cell removal the tissues were processed were incubated in DMEM 10% FBS containing Collagenase XI (300 units/ml, Sigma, C7657), Dispase II (0.1 mg/ml, Sigma, D4693) and DNase II Type V (50 units/ml, Sigma, D8764) for 1 h, at 37 °C, 200 rpm. Cells released after vigorous shaking were passed through a 70 µm strainer and washed with 2% sorbitol.

For organoid samples organoids were dissociated into single cell suspensions by incubation at 37 °C in 0.25% trypsin-EDTA solution (Gibco, 25200056) diluted 1:1 with DMEM without serum.

Instrument

FACS-sorting was performed at the Yale Flow Cytometry Facility with a BD FACSAria II sorter equipped with FACSDiva 7 software and with a BD FACSAria III sorter equipped with FACSDiva software at the Flow Cytometry Facility of BSRC Fleming.

Software

Data were acquired with the BD FACSDiva 7 software and analyzed with the FlowJo V10 software.

Cell population abundance

All sorted cells per population were used for RNA isolation

Gating strategy

Extended Data Figure 3c: FSC-A, SSC-A live cells >> SSC-W, SSC-H singlets >> FSC-H, FSC-W singlets >> Cd45-Tomato+, Cd45-Tomato-
 Extended Data Figure 3d: FSC-A, SSC-A live cells >> SSC-W, SSC-H singlets >> FSC-H, FSC-W singlets >> Cd45->>Tomato-GFP+
 Extended Data Figure 3b: FSC-A, FSC-H singlets >> FSC-A, SSC-A live cells >> Cd45-Pacific blue-, Pdgfra-EGFP+ >> Col6Cre tdTomato
 Extended Data Figure 7b: FSC-A, FSC-H singlets >> FSC-A, SSC-A live cells >> FSC, Cd24+ >> FSC, Sca-1

☒ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

The gluconeogenic enzyme PCK1 phosphorylates INSIG1/2 for lipogenesis

<https://doi.org/10.1038/s41586-020-2183-2>

Received: 7 December 2018

Accepted: 6 February 2020

Published online: 8 April 2020

 Check for updates

Daqian Xu^{1,2,16}, Zheng Wang^{3,16}, Yan Xia^{4,5}, Fei Shao⁶, Weiya Xia², Yongkun Wei², Xinjian Li⁷, Xu Qian⁸, Jong-Ho Lee⁹, Linyong Du¹⁰, Yanhua Zheng⁴, Guishuai Lv¹¹, Jia-shiun Leu¹², Hongyang Wang¹¹, Dongming Xing^{6,13}, Tingbo Liang¹, Mien-Chie Hung¹⁴ & Zhimin Lu^{6,15}

Cancer cells increase lipogenesis for their proliferation and the activation of sterol regulatory element-binding proteins (SREBPs) has a central role in this process. SREBPs are inhibited by a complex composed of INSIG proteins, SREBP cleavage-activating protein (SCAP) and sterols in the endoplasmic reticulum. Regulation of the interaction between INSIG proteins and SCAP by sterol levels is critical for the dissociation of the SCAP–SREBP complex from the endoplasmic reticulum and the activation of SREBPs^{1,2}. However, whether this protein interaction is regulated by a mechanism other than the abundance of sterol—and in particular, whether oncogenic signalling has a role—is unclear. Here we show that activated AKT in human hepatocellular carcinoma (HCC) cells phosphorylates cytosolic phosphoenolpyruvate carboxykinase 1 (PCK1), the rate-limiting enzyme in gluconeogenesis, at Ser90. Phosphorylated PCK1 translocates to the endoplasmic reticulum, where it uses GTP as a phosphate donor to phosphorylate INSIG1 at Ser207 and INSIG2 at Ser151. This phosphorylation reduces the binding of sterols to INSIG1 and INSIG2 and disrupts the interaction between INSIG proteins and SCAP, leading to the translocation of the SCAP–SREBP complex to the Golgi apparatus, the activation of SREBP proteins (SREBP1 or SREBP2) and the transcription of downstream lipogenesis-related genes, proliferation of tumour cells, and tumorigenesis in mice. In addition, phosphorylation of PCK1 at Ser90, INSIG1 at Ser207 and INSIG2 at Ser151 is not only positively correlated with the nuclear accumulation of SREBP1 in samples from patients with HCC, but also associated with poor HCC prognosis. Our findings highlight the importance of the protein kinase activity of PCK1 in the activation of SREBPs, lipogenesis and the development of HCC.

INSIG proteins are anchor proteins of the endoplasmic reticulum (ER) that have two isoforms, INSIG1 and INSIG2^{3,4}. The binding of cholesterol-derived oxysterols, including 22-, 24-, 25- and 27-hydroxycholesterol, is crucial for the binding of INSIG proteins to SREBP cleavage-activating protein (SCAP), and for the retention of the SREBP–SCAP complex in the ER^{5,6}. Under sterol-limiting conditions, the SREBP–SCAP complex is captured by COPII-coated vesicles and transported to the Golgi apparatus, where SIP and S2P proteases cleave SREBPs (SREBP1 or SREBP2) to yield active amino-terminal fragments for nuclear translocation and gene transcription¹. In

cancer cells, whether the interaction between INSIG proteins and the SREBP–SCAP complex is regulated by a mechanism that is independent of sterol abundance is unclear.

AKT-phosphorylated PCK1 binds to INSIG1/2

To investigate the regulation of INSIG proteins by oncogenic signalling, we treated Huh7 human HCC cells with insulin-like growth factor 1 (IGF1) for 1 h to induce the signalling that is critical for HCC development⁷. Mass spectrometric analyses of immunoprecipitates of

¹Department of Hepatobiliary and Pancreatic Surgery and Zhejiang Provincial Key Laboratory of Pancreatic Disease of The First Affiliated Hospital, Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, China. ²Department of Molecular and Cellular Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ³The Brown Foundation Institute of Molecular Medicine, The University of Texas Health Science Center at Houston, Houston, TX, USA. ⁴Department of Neuro-Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ⁵Department of Cancer Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. ⁶The Affiliated Hospital of Qingdao University and Qingdao Cancer Institute, Qingdao, China. ⁷CAS Key Laboratory of Infection and Immunity, CAS Center for Excellence in Biomacromolecules, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China. ⁸Jiangsu Key Laboratory of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Personalized Medicine, School of Public Health, Nanjing Medical University, Nanjing, China. ⁹Department of Biological Sciences, Dong-A University, Busan, South Korea. ¹⁰Key Laboratory of Laboratory Medicine, Ministry of Education of China, School of Laboratory Medicine and Life Science, Wenzhou Medical University, Wenzhou, China. ¹¹International Co-operation Laboratory on Signal Transduction, Eastern Hepatobiliary Surgery Institute, Second Military Medical University, Shanghai, China. ¹²Department of Neurosurgery, Houston Methodist Research Institute, Houston, TX, USA. ¹³School of Life Sciences, Tsinghua University, Beijing, China. ¹⁴Graduate Institute of Biomedical Sciences and Center for Molecular Medicine, and Office of the President, China Medical University, Taichung, Taiwan. ¹⁵Zhejiang Provincial Key Laboratory of Pancreatic Disease of The First Affiliated Hospital, Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, China. ¹⁶These authors contributed equally: Daqian Xu, Zheng Wang. ✉e-mail: xudaqian@zju.edu.cn; mhung@mail.cmu.edu.tw; zhiminlu@zju.edu.cn

INSIG1 and INSIG2 (Extended Data Fig. 1a, Supplementary Tables 1, 2) showed that IGF1 induced an association between INSIG1 or INSIG2 (hereafter, INSIG1/2) and cytosolic PCK1 (also known as PEPCK1). PCK1 is the rate-limiting enzyme of gluconeogenesis in the liver and kidney, and converts oxaloacetate and GTP into phosphoenolpyruvate and CO₂⁸. In humans, cytosolic PCK1 shares 63.4% sequence identity with mitochondrial PCK2⁹. Co-immunoprecipitation and immunofluorescence analyses showed that PCK1, but not PCK2, bound to INSIG1/2 in IGF1-stimulated Huh7 and Hep3B HCC cells (Fig. 1a, Extended Data Fig. 1b, c), and that IGF1 induced the colocalization of PCK1, but not PCK2, with INSIG1 (Extended Data Fig. 1d). In addition, cell fractionation analyses showed that a small amount of PCK1, but not PCK2, translocated to the ER (Extended Data Fig. 1e, f). Thus, PCK1 translocates to the ER and binds to INSIG1/2 upon IGF1 stimulation.

To determine the mechanism that regulates the interaction between PCK1 and INSIG1/2, we inhibited the signalling pathway downstream of IGF1 in Huh7 cells (Extended Data Fig. 1g). We found that treatment with the AKT inhibitor MK-2206 or expression of a dominant-negative AKT mutant (AKT-DN) blocked the IGF1-induced binding of PCK1 to INSIG1/2 (Fig. 1b, Extended Data Fig. 1h), the translocation of PCK1 to the ER (Extended Data Fig. 1i, j) and its colocalization with INSIG1 (Extended Data Fig. 1k). By contrast, expression of a constitutively active form of AKT (myr-AKT) induced the binding of PCK1, but not PCK2, to INSIG1/2 (Extended Data Fig. 1l). These results indicate that AKT induces the translocation of PCK1 to the ER, where it binds to INSIG1/2.

Co-immunoprecipitation analyses showed that stimulating HCC cells with IGF1 induced an interaction between AKT and PCK1 (Extended Data Fig. 1m). We also performed a pull-down assay that revealed that purified active glutathione S-transferase (GST)-tagged AKT1 bound directly to His-tagged PCK1 (Extended Data Fig. 2a) through its catalytic domain (as evidenced by the expression of different AKT1 truncation mutants; Extended Data Fig. 2b). An *in vitro* phosphorylation assay (Extended Data Fig. 2c) and liquid chromatography–tandem mass spectrometry (LC–MS/MS) analysis (Extended Data Fig. 2d) showed that active, but not inactive, AKT1 phosphorylated PCK1 at the evolutionally conserved residue Ser90 (Extended Data Fig. 2e) in an AKT phosphorylation motif RXXS/T¹⁰ (Extended Data Fig. 2f). Mutation of

Ser90 in PCK1 to alanine (S90A) (Fig. 1c) or treatment with MK-2206 (Extended Data Fig. 2g) abolished this phosphorylation, which was also detected using an antibody that is specific to PCK1 phosphorylated at Ser90 (PCK1(pS90)) (Extended Data Fig. 2c, h). Treatment with IGF1 rapidly induced the phosphorylation of PCK1 in Huh7 cells, and this was inhibited by pretreating cells with the AKT inhibitor MK-2206 (Fig. 1d). An RNA-interference-resistant (r) PCK mutant (rPCK1(S90A)) expressed in endogenous PCK1-depleted Huh7 cells was resistant to IGF1-induced phosphorylation of PCK1 Ser90 (Extended Data Fig. 2i) and did not exhibit myr-AKT1- or IGF1-induced translocation to the ER (Extended Data Fig. 2j, k). By contrast, the phosphorylation-mimicking PCK1(S90E) mutant accumulated in the ER without IGF1 stimulation (Extended Data Fig. 2k). Knock-in expression of PCK1(S90A) in HCC cells using CRISPR–Cas9-mediated genome editing (Extended Data Fig. 2l–n) blocked the IGF1-induced translocation of PCK1 to the ER and its colocalization with INSIG1 (Extended Data Fig. 2o, p). These results suggest that AKT1-mediated phosphorylation of PCK1 Ser90 is necessary and sufficient for the translocation of PCK1 to the ER.

Of note, AKT-mediated phosphorylation of PCK1 (Extended Data Fig. 2q–s) or the introduction of an S90E mutation (Extended Data Fig. 2t–v) reduced the binding affinity of PCK1 to oxaloacetate and its enzymatic activity (that is, the production of phosphoenolpyruvate). Thus, AKT-mediated phosphorylation of PCK1 and its translocation to the ER inhibit the canonical function of PCK1 in gluconeogenesis.

To determine the role of the phosphorylation of PCK1 Ser90 in its binding to INSIG1/2, we mixed purified INSIG1/2 with wild-type PCK1 or GST–PCK1 or GST–PCK1(S90A). Only AKT-phosphorylated wild-type PCK1 interacted with INSIG1/2 (Extended Data Fig. 2w). This interaction was abolished by treatment with calf intestinal alkaline phosphatase (CIP), which dephosphorylated PCK1(pS90) residue. Consistent with this, treatment with IGF1 induced the binding of INSIG1/2 to S protein–Flag–streptavidin-binding-peptide (SFB)-tagged wild-type PCK1, but not SFB–PCK1(S90A) (Extended Data Fig. 2x). This binding was abrogated by treatment with CIP (Extended Data Fig. 2x) or by knock-in expression of PCK1(S90A) in HCC cells (Fig. 1e, Extended Data Fig. 2y). Thus, phosphorylation of PCK1 Ser90 is required for the binding of PCK1 to INSIG1/2.

To determine the role of INSIG1/2 in the translocation of PCK1 to the ER, we depleted INSIG1/2 in Huh7 cells, and found that this blocked the translocation of PCK1 to the ER that was induced by IGF1 or by myr-AKT1 (Extended Data Fig. 2z). Expression of INSIG1/2 truncation mutants revealed that loop 1 of INSIG1/2 bound to PCK1 (Extended Data Fig. 3a). This interaction was not blocked in a PCK1(pS90) peptide (Extended Data Fig. 3b), which suggests that it is a conformational change in PCK1 (mediated by phosphorylation at Ser90) that causes the binding of PCK1 to INSIG1/2, rather than the phosphorylation of Ser90 directly.

PCK1 phosphorylates INSIG1/2

As PCK1 is able to transfer a phosphate group from GTP to a metabolite, we next investigated whether PCK1 phosphorylates INSIG1/2. In the presence of radiolabelled [γ -³²P]GTP and active AKT1, only purified wild-type PCK1 and not PCK1(S90A) or PCK1(C288S)—a kinase-dead mutant that could still be phosphorylated at Ser90 and interacted with INSIG1/2 (Extended Data Fig. 3c)—phosphorylated purified INSIG1 (Fig. 2a) and INSIG2 (Extended Data Fig. 3d). By contrast, this phosphorylation did not occur in the presence of [γ -³²P]ATP (Extended Data Fig. 3e). LC–MS/MS analyses showed that INSIG1 was phosphorylated at Ser207 (Extended Data Fig. 3f), which corresponds to Ser151 in INSIG2. Both of these residues are evolutionally conserved and are located in the cytosolic loop 2 of INSIG1/2 (Extended Data Fig. 3g, h). INSIG1(S207A) (Fig. 2b) and INSIG2(S151A) (Extended Data Fig. 3i) mutants were not phosphorylated by PCK1 *in vitro* (as detected by an antibody that specifically recognizes both phosphorylated residues; Extended Data Fig. 3j, k). Compared to wild-type PCK1, PCK1(S90E) exhibited a much

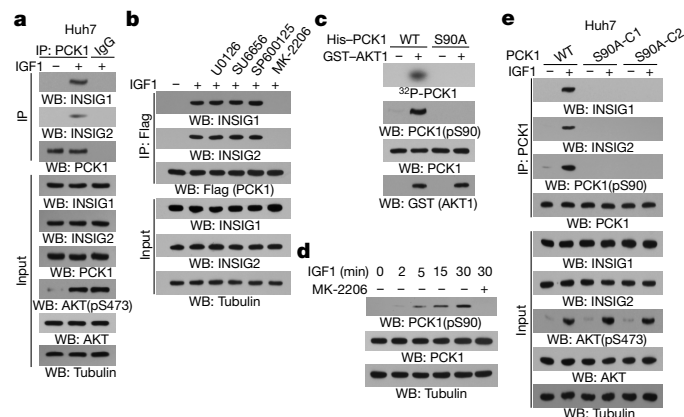


Fig. 1 | IGF1-induced and AKT-phosphorylated PCK1 translocates to the ER and binds to INSIG1/2. **a–e**, Immunoprecipitation (IP) and western blotting (WB) analyses were performed as indicated three times with similar results. **a**, Huh7 cells were treated with or without IGF1 for 1 h. **p** indicates a phosphorylated residue. **b**, Huh7 cells expressing Flag–PCK1 were pretreated with or without the indicated inhibitors for 30 min before treatment with or without IGF1 for 1 h. **c**, *In vitro* kinase assays were performed by mixing purified wild-type (WT) His–PCK1 or His–PCK1(S90A) with or without purified GST–AKT1 in the presence of [γ -³²P]ATP. **d**, Huh7 cells pretreated with or without MK-2206 for 30 min were treated with IGF1 for the indicated time periods. **e**, Parental Huh7 cells and the indicated clones with knock-in expression of PCK1(S90A) were stimulated with or without IGF1 for 1 h. C1, clone 1; C2, clone 2.

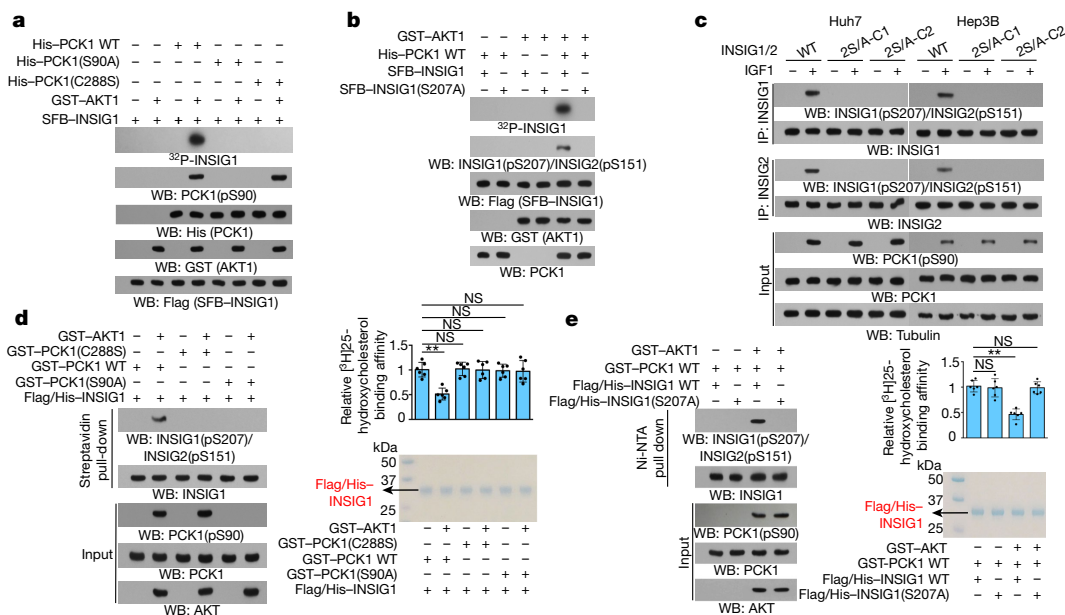


Fig. 2 | PCK1 phosphorylates INSIG1 Ser207 and INSIG2 Ser151, thereby reducing the binding of oxysterols to INSIG1/2. **a–e**, Immunoblotting analyses were performed as indicated three times with similar results. **a, b**, Bacterially purified His–PCK1 proteins on Ni–NTA agarose beads were incubated with or without active GST–AKT1 in the presence of ATP for an in vitro kinase assay. The beads were then washed and incubated with or without the indicated SFB–INSIG1 proteins in the presence of [γ - 32 P]GTP. Autoradiography was performed. **c**, Parental Huh7 and Hep3B cells and the indicated clones with knock-in expression of INSIG1(S207A)/INSIG2(S151A)

double mutants (2S/A) were stimulated with or without IGF1 for 1 h. **d, e**, Flag and His (Flag/His)-tagged wild-type INSIG1 (**d**) or INSIG1(S207A) (**e**) immunoprecipitated and purified from Huh7 cells were incubated with the indicated GST–PCK1 proteins with or without active GST–AKT1 in the presence of ATP and GTP for 1 h. The INSIG proteins on Ni–NTA agarose beads were washed and incubated with 400 nM [3 H]25-hydroxycholesterol. Specifically bound [3 H]25-hydroxycholesterol was measured (right) ($n = 6$). Data are mean \pm s.d. NS, not significant ($P = 0.859, 0.930, 0.795, 0.768$ (left to right) (**d**); $P = 0.720, 0.630$ (left to right) (**e**)); $^{**}P < 0.001$ (two-tailed t -test).

higher velocity of enzyme-catalysed reaction at infinite concentration of substrate (V_{\max}) and lower Michaelis constant (K_m) in phosphorylating an INSIG1 peptide at Ser207 (Extended Data Fig. 3l). Notably, PCK1(S90E) or PCK1(S90D) mutants phosphorylated Ser207 of INSIG1 and Ser151 of INSIG2 in the absence of AKT (Extended Data Fig. 3m, n). In addition, INSIG1(S207A) and INSIG2(S151A), expressed in Huh7 cells, were resistant to phosphorylation that was induced by IGF1 stimulation or myr-AKT1 expression (Extended Data Fig. 3o, p). In line with this, IGF1-induced phosphorylation of INSIG1/2 was also abolished by knock-in expression of PCK1(S90A) (Extended Data Fig. 3q), or by that of INSIG1(S207A)/INSIG2(S151A) (Fig. 2c, Extended Data Fig. 4a–d) in HCC cells. Thus, AKT-phosphorylated PCK1 functions as a protein kinase and uses GTP as a phosphate donor to phosphorylate INSIG1/2.

INSIG1/2 phosphorylation reduces sterol binding

We next examined whether phosphorylation of INSIG1/2 affects the binding of INSIG proteins to oxysterols. Phosphorylation of wild-type INSIG1 and INSIG2 by AKT-phosphorylated wild-type PCK1 (Fig. 2d, Extended Data Fig. 4e), PCK1(S90D) or PCK1(S90E) (Extended Data Fig. 4f, g)—but not PCK1(C288S) or PCK1(S90A)—reduced the binding affinity of INSIG1/2 to [3 H]25-hydroxycholesterol. By contrast, the binding affinity was unchanged for INSIG1(S207A) (Fig. 2e) and INSIG2(S151A) (Extended Data Fig. 4h). Similarly, wild-type INSIG1/2—but not INSIG1(S207A) or INSIG2(S151A) (Extended Data Fig. 4i, j)—that was immunoprecipitated from IGF1-treated parental Huh7 cells showed a reduction in binding to [3 H]25-hydroxycholesterol, whereas the binding affinity of wild-type INSIG1/2 that was immunoprecipitated from Huh7 cells with knock-in expression of PCK1(S90A) was not affected (Extended Data Fig. 4k, l). This reduction in binding also occurred for the phospho-mimicking mutants INSIG1(S207E) and INSIG2(S151E) (Extended Data Fig. 4m, n). These results indicate that phosphorylation of INSIG1/2 by PCK1 reduces the binding of INSIG1/2 to oxysterols.

The PCK1–INSIG1/2 axis activates SREBP

The release of oxysterols from INSIG1/2 results in the disruption of the interaction between INSIG1/2 and SCAP, the translocation of the SCAP–SREBP1 complex from the ER to the Golgi apparatus and nuclear accumulation of SREBP1⁵. As expected, stimulation with IGF1 disrupted the association between INSIG1/2 and SCAP (Fig. 3a, Extended Data Fig. 5a). This disruption was inhibited in Huh7 and Hep3B cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) (Fig. 3a, Extended Data Fig. 5a) or PCK1(S90A) (Extended Data Fig. 5b, c), or with reconstituted expression of rPCK1(S90A) (Extended Data Fig. 5d, e) or rPCK1(C288S) (Extended Data Fig. 5f). Cell fractionation analyses showed that IGF1 treatment induced the translocation of SCAP from the ER to the Golgi apparatus, and that this was inhibited by knock-in expression of INSIG1(S207A)/INSIG2(S151A) (Extended Data Fig. 5g, h) or PCK1(S90A) (Extended Data Fig. 5i, j). Expression of these mutants also blocked the IGF-induced loss of colocalization of SCAP with the ER protein calnexin (Fig. 3b, Extended Data Fig. 5k, n) and the colocalization of SCAP with the Golgi apparatus protein golgin-97 (Extended Data Fig. 5l, m, o). Similar results were obtained when we expressed Myc-tagged SCAP (Extended Data Fig. 5p, q).

Stimulation with IGF1 increased the cleavage of SREBP1 (Extended Data Figs. 5r, 6a) and its nuclear accumulation (Fig. 3c, Extended Data Fig. 6b); SRE-promoter-driven luciferase activity (Extended Data Figs. 5s, 6c); and the expression of mRNA and proteins of SREBP1 target genes that are associated with lipogenesis, including fatty acid synthase (*FASN*), acetyl-CoA carboxylase alpha (*ACACA*, also known as *ACC1*), stearoyl-CoA desaturase (*SCD*, also known as *SCD1*) and glycerol-3-phosphate acyltransferase 1 (*GPAM*, also known as *GPAT*), as well as *SREBF1* (which encodes SREBP1)^{1,11,12} (Fig. 3d, Extended Data Figs. 5t, 6d–f). Consequently, treatment with IGF1 resulted in increased incorporation of 14 C-glucose into triglycerides and fatty acids (Fig. 3e, Extended Data Fig. 6g). Notably, all of these IGF1-induced

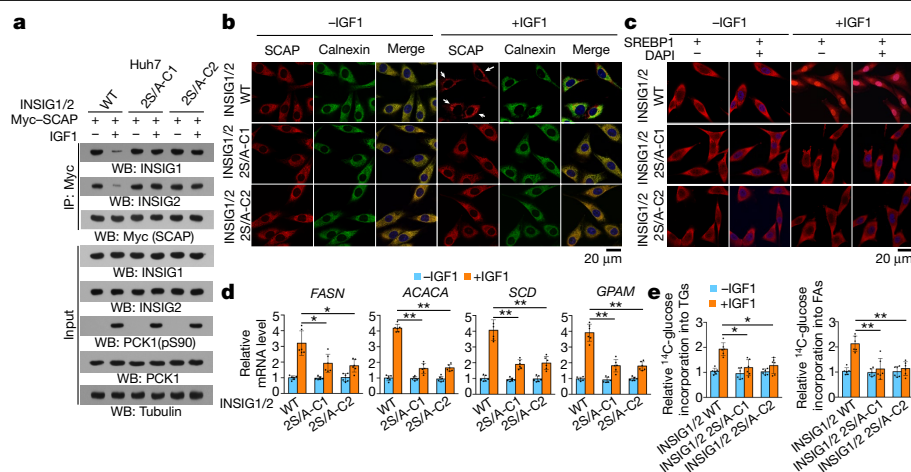


Fig. 3 | PCK1-mediated phosphorylation of INSIG1/2 releases SCAP from the ER and promotes SREBP1 activation for lipogenesis. **a–c**, Parental Huh7 cells and INSIG1(S207A)/INSIG2(S151A) knock-in cells were transfected with the indicated plasmids and stimulated with or without IGF1 for 16 h. Immunoprecipitation and immunoblotting analyses (a) and immunofluorescence analyses (b, c) were performed as indicated. The arrows in b indicate Golgi-localized SCAP. The experiments were repeated three times

independently with similar results. **d**, The mRNA expression levels of SREBP1 target genes were measured using quantitative PCR. * P = 0.009, 0.002 (left to right); ** P < 0.001 (two-tailed t -test). **e**, The incorporation of 14 C-glucose into triglycerides (TGs) (left) and fatty acids (FAs) (right) was measured. * P = 0.01, 0.03 (left to right); ** P < 0.001 (two-tailed t -test). Data in **d**, **e** are mean \pm s.d. (n = 6 biological replicates).

effects were inhibited by knock-in expression of INSIG1(S207A)/INSIG2(S151A) (Fig. 3c–e, Extended Data Figs. 5r–t, 6f) or PCK1(S90A) (Extended Data Fig. 6a–g). Furthermore, the IGF1-induced incorporation of 14 C-glucose into fatty acids was strongly blocked by expression of PCK1(S90A) in CHL-1 melanoma cells, U87 glioblastoma cells and H1993 non-small-cell lung cancer cells (Extended Data Fig. 6h), suggesting that the lipogenesis that is promoted by phosphorylation of PCK1 Ser90 occurs in several types of cancer. Consistently, knock-in expression of PCK1(S90D) (Extended Data Figs. 2l–n, 6i) or INSIG1(S207D)/INSIG2(S151D) (Extended Data Figs. 4a–d, 6j) in HCC cells was sufficient to induce SREBP1 cleavage.

Similar to our observations with SREBP1, expression of PCK1(S90A) or INSIG1(S207A)/INSIG2(S151A) also inhibited IGF1-induced cleavage of SREBP2 (Extended Data Fig. 6k, l) and SREBP2-mediated transcription of genes that are related to cholesterol biogenesis¹³, such as 3-hydroxy-3-methylglutaryl-CoA (HMGCoA) reductase (*HMGCR*), HMGCoA synthase 1 (*HMGCS1*, also known as *HMGCS*), low-density lipoprotein receptor (*LDLR*) and squalene synthase (*FDFT1*, also known as *SS*) (Extended Data Fig. 6m, n). Thus, PCK1-mediated phosphorylation of INSIG1/2 promotes the activation of both SREBP1 and SREBP2, and the expression of downstream genes that are involved in lipogenesis.

Because PCK1-mediated phosphorylation of INSIG1/2 promotes the release of sterols from their binding to INSIG1/2, we postulated that this phosphorylation does not affect the activation of SREBP1 that is induced by lipid depletion. As expected, knock-in expression of INSIG1(S207A)/INSIG2(S151A) or PCK1(S90A) did not affect the lipid-depletion-induced translocation of SCAP from the ER to the Golgi apparatus (Extended Data Fig. 6o–r), loss of colocalization of SCAP with calnexin (Extended Data Fig. 6s–u) and colocalization of SCAP with golgin-97 (Extended Data Fig. 6v–x), the cleavage and nuclear accumulation of SREBP1 (Extended Data Fig. 7a, b), and SRE-driven luciferase activity (Extended Data Fig. 7c). By contrast, incubation of HCC cells with 25-hydroxycholesterol—but not cholesterol—at a dosage (120 nM) much higher than physiological concentrations (40 nM)¹⁴ blocked the dissociation of INSIG1/2 from SCAP, the cleavage of SREBP and the increase of SRE luciferase activity that were induced by IGF1 (Extended Data Fig. 7d, f) or by expression of phosphorylation-mimicking PCK1 and INSIG1/2 mutants (Extended Data Fig. 7e, g). These results demonstrate that PCK1-mediated phosphorylation of INSIG1/2 promotes

SREBP1 activation through the release of 25-hydroxycholesterol from INSIG1/2.

SREBP1 can be activated both in a manner that depends on AKT and mTORC1, and independently of AKT and mTORC1^{15–17}. It has been previously shown that insulin signalling and AKT activation suppress the expression of INSIG2 by promoting the decay of its mRNA¹⁷. In addition, INSIG1, but not INSIG2, is ubiquitinated and degraded upon sterol depletion¹⁸. A time-course experiment showed that INSIG2 was not obviously downregulated until prolonged treatment with IGF1 (24–48 h) (Extended Data Fig. 7h), and this downregulation was not affected in PCK1(S90A) or INSIG1(S207A)/INSIG2(S151A) mutants (Extended Data Fig. 7i). By contrast, treatment of HCC cells with IGF1 and cycloheximide (CHX)—which eliminates the translational regulation of protein expression—induced rapid degradation of wild-type INSIG1, but not INSIG1(S207A), wild-type INSIG2 or INSIG2(S151A) (Extended Data Fig. 7j, k). In addition, the half-life of INSIG1(S207D), but not INSIG2(S151D), was shortened (Extended Data Fig. 7l, m), indicating that phosphorylation of INSIG1 Ser207 promotes the degradation of the INSIG1 protein. However, the overall expression of INSIG1 in the presence of IGF1 and absence of CHX was not obviously altered (Extended Data Fig. 7h). This might be a result of an increase in *INSIG1* transcription owing to SREBP activation¹⁹, which compensates for the degradation of INSIG1 in a feedback manner. In contrast to the delayed response of INSIG2 expression to IGF1 stimulation, immediate phosphorylation of PCK1 and INSIG1/2 (Extended Data Fig. 7n) and rapid cleavage of SREBP1 (Extended Data Fig. 7h) were detected after IGF1 treatment. Thus, PCK1-mediated rapid phosphorylation of INSIG1/2 and activation of SREBP1 is an immediate response to AKT activation.

De novo fatty acid synthesis occurs in the liver, adipose tissue and lactating breast²⁰. Treatment of HL7702 and THLE-2 normal hepatocytes with IGF1 induced phosphorylation of PCK1 Ser90, subsequent phosphorylation of INSIG1 Ser207 and INSIG2 Ser151, and SREBP1 cleavage (Extended Data Fig. 8a). The latter effect was also induced by expression of PCK1(S90D) in these cells (Extended Data Fig. 8b), which suggests that PCK1-induced activation of SREBP1 also occurs in normal hepatocytes.

Under normal physiological conditions, high blood glucose levels after a meal increase the pancreatic secretion of insulin, which immediately activates the phosphoinositide 3-kinase (PI3K)–AKT signalling

pathway and leads to an increase in glucose utilization and a reduction in gluconeogenesis in the liver²¹. Of note, phosphorylation of AKT, PCK1 Ser90, INSIG1 Ser207 and INSIG2 Ser151, and cleavage of SREBP1, were markedly enhanced in normal liver from mice that were refed with glucose after 24 h of fasting (Extended Data Fig. 8c)—suggesting that *in vivo* blood glucose levels regulate the PCK1-mediated phosphorylation of INSIG1/2 and the activation of SREBP1 in the liver. These results demonstrate the relevance of our findings in the context of the physiological functions of the liver.

We then examined a potential difference in the PCK1-mediated activation of SREBP1 between normal hepatocytes and HCC cells. We found that there was a substantial increase in the phosphorylation of AKT, PCK1 Ser90, INSIG1 Ser207 and INSIG2 Ser151—and in the cleavage of SREBP1—in HCC cells compared with normal human hepatocytes (Extended Data Fig. 8d). These increases were reduced by expression of PCK1(S90A) in HCC cells (Extended Data Fig. 8e), indicating that PCK1-mediated SREBP1 activation is increased in HCC cells with a high level of AKT activation.

Notably, these increases in the phosphorylation of AKT, PCK1 and INSIG1/2, and in the cleavage of SREBP1, were further enhanced or induced in Huh7, CHL-1, U87 and H1993 cancer cells that express the active KRAS(G12V) mutant (Extended Data Fig. 8f, g), a mutant version of the IGF1 receptor (IGF1R) (IGF1R(V922E)) (Extended Data Fig. 8h) or a mutant version of the epidermal growth factor receptor (EGFR) (EGFRvIII) (Extended Data Fig. 8i), and in cells treated with platelet-derived growth factor (PDGF) (Extended Data Fig. 8j). Expression of PCK1(S90A) (Extended Data Fig. 8f–j) or INSIG1(S207A)/INSIG2(S151A) (Extended Data Fig. 8f, h) inhibited INSIG1/2 phosphorylation and SREBP1 cleavage in these cells. These results suggest that PCK1-mediated activation of SREBP1 is induced by AKT activation that is elicited by different oncogenes or growth factors and occurs in several types of cancer.

The PCK1–INSIG1/2 axis promotes HCC growth

Notably, reconstituted expression of PCK1(S90A) in different HCC cell lines—which reduced the basal level of phosphorylation of INSIG1 Ser207 and INSIG2 Ser151 and the cleavage of SREBP1 (Extended Data Fig. 8k)—inhibited the proliferation of cells under normal culture conditions (Extended Data Fig. 8m), without altering the cleavage of SREBP1 and the decrease in cell survival induced by lipid depletion (Extended Data Fig. 8l, n). Similarly, knock-in expression of INSIG1(S207A)/INSIG2(S151A) or PCK1(S90A) inhibited the proliferation of Huh7 cells (Fig. 4a) and Hep3B cells (Extended Data Fig. 9a). By contrast, expression of PCK1(S90D) or INSIG1(S207D)/INSIG2(S151D) mutants enhanced cell proliferation (Extended Data Fig. 9b, c).

Next, we intrahepatically (Fig. 4b) and subcutaneously (Extended Data Fig. 9d–f) injected Huh7 cells with or without knock-in expression of PCK1(S90A) or INSIG1(S207A)/INSIG2(S151A) into nude mice. Expression of these mutant proteins substantially inhibited tumour growth in the mice (Fig. 4b, Extended Data Fig. 9d–f), with a corresponding reduction in Ki67 expression (Extended Data Fig. 9g) and an increase in cell apoptosis (Extended Data Fig. 9h). In addition, expression of PCK1(S90A) reduced the phosphorylation of INSIG1 Ser207 and INSIG2 Ser151, and reduced the expression and nuclear distribution of SREBP1 (Fig. 4c, Extended Data Fig. 9i, j, left). In line with this, expression of INSIG1(S207A)/INSIG2(S151A) also decreased the nuclear levels of SREBP1 (Fig. 4d, Extended Data Fig. 9i, j, right). By contrast, expression of PCK1(S90D) or INSIG1(S207D)/INSIG2(S151D) promoted tumour growth (Extended Data Fig. 9k–m).

Consistent with the finding that AKT–PCK1 signalling activates SREBP1, expression of active IGF1R(V922E) (Extended Data Fig. 10a–c) or myr-AKT (Extended Data Fig. 10d–f) in Huh7 cells significantly increased the growth of liver tumours, with enhanced phosphorylation of PCK1 Ser90 and INSIG1 Ser207/INSIG2 Ser151 and enhanced

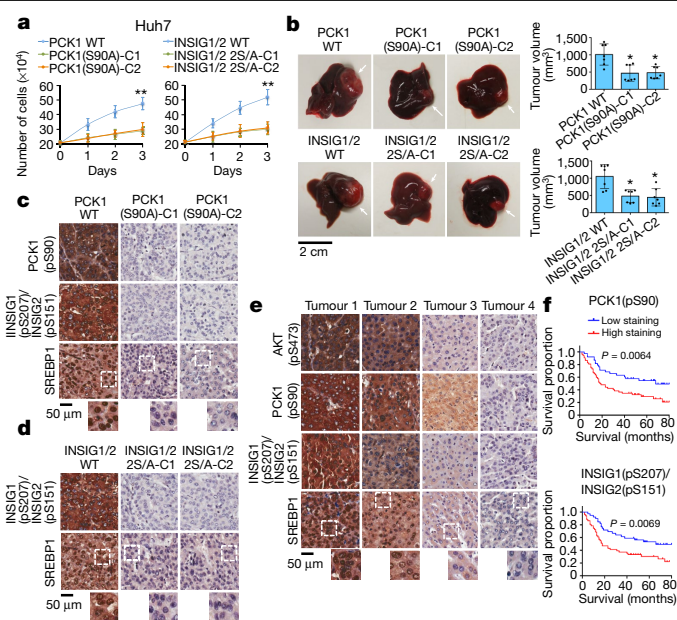


Fig. 4 | PCK1-mediated phosphorylation of INSIG1/2 promotes liver tumour growth and correlates with poor prognosis for HCC. **a**, Parental Huh7 cells (2×10^5) and the indicated clones with knock-in expression of PCK1(S90A) (left) or INSIG1(S207A)/INSIG2(S151A) (right) were plated for 3 days. The cells were then collected and counted. ** $P < 0.001$ (two-tailed *t*-test). **b**, Huh7 cells with or without knock-in expression of PCK1(S90A) (top) or INSIG1(S207A)/INSIG2(S151A) (bottom) were intrahepatically injected into athymic nude mice ($n = 7$ per group). Tumour growth was examined 28 days after injection. The arrows indicate tumours. Tumour volumes were calculated (right). * $P = 0.03$, 0.02 (left to right, top); * $P = 0.02$, 0.03 (left to right, bottom) (two-tailed *t*-test). Data in **a**, **b** are mean \pm s.d. **c**–**e**, IHC analyses of xenograft tumours from nude mice ($n = 7$) (**c**, **d**) and 90 human HCC samples (**e**) were performed with the indicated antibodies. Representative staining images are shown. The regions in white boxes are shown at higher magnification below. **f**, Kaplan–Meier plots of the overall survival rates in 90 patients with HCC grouped according to high (staining score, 4–8) and low (staining score, 0–3) expression of PCK1(pS90) (top) and INSIG1(pS207)/INSIG2(pS151) (bottom) ($n = 38$ (PCK1(pS90) low), $n = 52$ (PCK1(pS90) high); $n = 39$ (INSIG1(pS207)/INSIG2(pS151) low), $n = 51$ (INSIG1(pS207)/INSIG2(pS151) high). *P* values were calculated using a log-rank test (two-tailed).

nuclear expression of SREBP1. Knock-in expression of PCK1(S90A) or INSIG1(S207A)/INSIG2(S151A) inhibited both basal tumour growth and that induced by IGF1R(V922E) (Extended Data Fig. 10a–c) or myr-AKT (Extended Data Fig. 10d–f). By contrast, expression of a dominant-negative IGF1R(L1003R) mutant reduced both intrahepatic (Extended Data Fig. 10g, h) and subcutaneous (Extended Data Fig. 10i–n) tumour growth, with reduced phosphorylation of PCK1 and INSIG1/2 (Extended Data Fig. 10o) and reduced nuclear accumulation of SREBP1 (Extended Data Fig. 10p, q). The reduction in tumour growth (Extended Data Fig. 10g–n) and nuclear SREBP1 expression (Extended Data Fig. 10p, q) was partially reverted by expression of PCK1(S90D) or INSIG1(S207D)/INSIG2(S151D).

Furthermore, we used hydrodynamics-based transfection in mice to administer plasmids for the expression of active myr-AKT, c-Met and the sleeping beauty transposase (which induces rapid liver tumour growth²²) together with wild-type PCK1 or PCK1(S90A). Expression of PCK1(S90A) reduced tumour growth, Ki67 expression, INSIG1/2 phosphorylation and nuclear SREBP1 expression (Extended Data Fig. 10r–t). Thus, PCK1-mediated phosphorylation of INSIG1/2 and subsequent activation of SREBP1 promote the development of HCC.

To determine the clinical relevance of PCK1-regulated SREBP1 activation, we performed immunohistochemistry (IHC) analyses of 30 paired

samples of primary HCC and adjacent normal tissue. Phosphorylation of PCK1 Ser90 and INSIG1 Ser207/INSIG2 Ser151—and nuclear SREBP1 expression—were markedly increased in the HCC specimens compared with normal tissue (Extended Data Fig. 10u–w), and correlated with each other in 90 resected HCC tumours (Fig. 4e, Extended Data Fig. 10x, y). Notably, in HCC samples, high levels of phosphorylation of PCK1 Ser90 and INSIG1 Ser207/INSIG2 Ser151—and high levels of nuclear SREBP1 expression—were correlated with decreased overall durations of survival in patients with HCC (Fig. 4f). These results suggest that PCK1-mediated phosphorylation of INSIG1/2 has a critical role in the clinical aggressiveness of human HCC.

Metabolism and gene expression are two fundamental cellular processes that are essential for the proliferation of tumour cells and that can be mutually regulated²³. PCK1 was originally characterized as a gluconeogenesis enzyme. Herein we have shown that PCK1 has protein kinase activity and translocates to the ER to regulate SREBP1 activation and SREBP1-mediated gene expression (Extended Data Fig. 10z). We also demonstrated that a metabolic enzyme uses GTP, rather than ATP, as a phosphate donor to phosphorylate a protein substrate. In addition, we have shown that oncogenic signalling can rapidly modulate the association between INSIG1/2 and sterol without reducing total cellular sterol levels. Our findings highlight the potential for inhibition of the protein kinase activity of PCK1 as a treatment strategy in human HCC.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2183-2>.

1. Goldstein, J. L., DeBose-Boyd, R. A. & Brown, M. S. Protein sensors for membrane sterols. *Cell* **124**, 35–46 (2006).
2. Röhrig, F. & Schulze, A. The multifaceted roles of fatty acid synthesis in cancer. *Nat. Rev. Cancer* **16**, 732–749 (2016).
3. Feramisco, J. D., Goldstein, J. L. & Brown, M. S. Membrane topology of human Insig-1, a protein regulator of lipid synthesis. *J. Biol. Chem.* **279**, 8487–8496 (2004).
4. Yabe, D., Brown, M. S. & Goldstein, J. L. Insig-2, a second endoplasmic reticulum protein that binds SCAP and blocks export of sterol regulatory element-binding proteins. *Proc. Natl Acad. Sci. USA* **99**, 12753–12758 (2002).

5. Radhakrishnan, A., Ikeda, Y., Kwon, H. J., Brown, M. S. & Goldstein, J. L. Sterol-regulated transport of SREBPs from endoplasmic reticulum to Golgi: oxysterols block transport by binding to Insig. *Proc. Natl Acad. Sci. USA* **104**, 6511–6518 (2007).
6. Shimano, H. & Sato, R. SREBP-regulated lipid metabolism: convergent physiology - divergent pathophysiology. *Nat. Rev. Endocrinol.* **13**, 710–730 (2017).
7. Breuhahn, K., Longerich, T. & Schirmacher, P. Dysregulation of growth factor signaling in human hepatocellular carcinoma. *Oncogene* **25**, 3787–3800 (2006).
8. Burgess, S. C. et al. Cytosolic phosphoenolpyruvate carboxykinase does not solely control the rate of hepatic gluconeogenesis in the intact mouse liver. *Cell Metab.* **5**, 313–320 (2007).
9. Méndez-Lucas, A., Hyroššová, P., Novellasdemunt, L., Viñals, F. & Perales, J. C. Mitochondrial phosphoenolpyruvate carboxykinase (PEPCK-M) is a pro-survival, endoplasmic reticulum (ER) stress response gene involved in tumor cell adaptation to nutrient availability. *J. Biol. Chem.* **289**, 22090–22102 (2014).
10. Vinayagam, A. et al. An integrative analysis of the InR/PI3K/Akt network identifies the dynamic response to insulin signaling. *Cell Rep.* **16**, 3062–3074 (2016).
11. Espenshade, P. J. SREBPs: sterol-regulated transcription factors. *J. Cell Sci.* **119**, 973–976 (2006).
12. Dif, N. et al. Insulin activates human sterol-regulatory-element-binding protein-1c (SREBP-1c) promoter through SRE motifs. *Biochem. J.* **400**, 179–188 (2006).
13. Horton, J. D., Goldstein, J. L. & Brown, M. S. SREBPs: activators of the complete program of cholesterol and fatty acid synthesis in the liver. *J. Clin. Invest.* **109**, 1125–1131 (2002).
14. Bakos, J. T., Johnson, B. H. & Thompson, E. B. Oxysterol-induced cell death in human leukemic T-cells correlates with oxysterol binding protein occupancy and is independent of glucocorticoid-induced apoptosis. *J. Steroid Biochem. Mol. Biol.* **46**, 415–426 (1993).
15. Guo, D. et al. EGFR signaling through an Akt-SREBP-1-dependent, rapamycin-resistant pathway sensitizes glioblastomas to antilipogenic therapy. *Sci. Signal.* **2**, ra82 (2009).
16. Düvel, K. et al. Activation of a metabolic gene regulatory network downstream of mTOR complex 1. *Mol. Cell* **39**, 171–183 (2010).
17. Yecies, J. L. et al. Akt stimulates hepatic SREBP1c and lipogenesis through parallel mTORC1-dependent and independent pathways. *Cell Metab.* **14**, 21–32 (2011).
18. Lee, J. N., Song, B., DeBose-Boyd, R. A. & Ye, J. Sterol-regulated degradation of Insig-1 mediated by the membrane-bound ubiquitin ligase gp78. *J. Biol. Chem.* **281**, 39308–39315 (2006).
19. Shao, W. & Espenshade, P. J. Expanding roles for SREBP in metabolism. *Cell Metab.* **16**, 414–419 (2012).
20. Menendez, J. A. & Lupu, R. Fatty acid synthase and the lipogenic phenotype in cancer pathogenesis. *Nat. Rev. Cancer* **7**, 763–777 (2007).
21. Huang, X., Liu, G., Guo, J. & Su, Z. The PI3K/AKT pathway in obesity and type 2 diabetes. *Int. J. Biol. Sci.* **14**, 1483–1496 (2018).
22. Hu, J. et al. Co-activation of AKT and c-Met triggers rapid hepatocellular carcinoma development via the mTORC1/FASN pathway in mice. *Sci. Rep.* **6**, 20484 (2016).
23. Li, X., Egervari, G., Wang, Y., Berger, S. L. & Lu, Z. Regulation of chromatin and gene expression by metabolic enzymes and metabolites. *Nat. Rev. Mol. Cell Biol.* **19**, 563–578 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Materials

Normal mouse IgG (sc-2025), normal rabbit IgG (sc-2027), GST (sc-138), tubulin (sc-8035), ERK1/2 (sc-514302) and INSIG1 (A-9) (sc-390504) (for immunoprecipitation or immunoblotting) antibodies were obtained from Santa Cruz Biotechnology. PCK1 (16754-1-AP) (for immunoprecipitation or immunoblotting), INSIG1 (22115-1-AP) (for immunofluorescence or immunoblotting) and INSIG2 (24766-1-AP) (for immunoprecipitation or immunoblotting) antibodies were purchased from Proteintech. Anti-AKT(pS473) (4060), rabbit AKT (9272), mouse AKT (2920), c-SRC (2108), haemagglutinin (HA) (3724), p44/42 MAPK (ERK1/2) (9102), c-Jun (9165), c-Jun(pS73) (3270), Myc-tag (9B11) (2276) and PCK1 (D12F5) (12940) antibodies (for immunoblotting) were purchased from Cell Signaling Technology. Rabbit antibodies that recognize PCK1(pS90), INSIG1(pS207) and INSIG2(pS151) were obtained from Signalway Biotechnology. IGF1 (85580C), PDGF (P8147), cyclohexamide (CHX) (01810), oxaloacetate (O7753), cholesterol (C8667), 25-hydroxycholesterol (H1015), mouse monoclonal anti-Flag (F1804), rabbit anti-Flag (F7425) and anti-His (SAB1305538) antibodies were purchased from Sigma-Aldrich. Anti-Flag M2 agarose beads were purchased from MP Biochemicals. [^3H]25-hydroxycholesterol, [γ - ^{32}P]ATP and [γ - ^{32}P]GTP were obtained from PerkinElmer. Inactive AKT1 protein (14-279-D) and rabbit anti-Ki67 antibody (AB9260) were obtained from Millipore. Oxaloacetate assay kit (ab83428), anti-Myc tag rabbit (ab9106), calnexin (ab22595), SRC(pY418) (ab4816), SCAP (ab91323), golgin-97 (ab84340), SCD1 (CD.E10) (ab19862), FASN (ab22759) and anti-rabbit IgG heavy chain (HRP) (ab99702) antibodies were purchased from Abcam. U0126, SP600125, MK-2206 and SU6656 were purchased from EMD Biosciences. Active GST-AKT1 (A16-10G) was obtained from SignalChem. Active recombinant His-AKT1 protein (LS-G18427) was obtained from Lifespan Biosciences. SREBP1 (IgG 2A4) and SREBP2 antibodies (557037) for immunoblotting analyses were purchased from BD Biosciences. SREBP1 antibody (2A4) (NB100-2215) (for immunofluorescence and IHC analyses) and PCK1 mouse antibody (3E4) (H00005105-M1) (for immunofluorescence or immunoblotting) were purchased from Novus. INSIG2 (PA5-41707) and INSIG1 (PA5-97876) antibodies (for immunoblotting), calnexin (AF18) (MA3-027), golgin-97 (CDF4) (A-21270), glutathione agarose, 4', 6-diamidino-2-phenylindole (DAPI), Alexa Fluor 488 goat anti-rabbit (A11008), Alexa Fluor 594 goat anti-rabbit (A11012), Alexa Fluor 488 goat anti-mouse (A11029) and Alexa Fluor 594 goat anti-mouse antibodies (A11005) were obtained from Thermo Fisher Scientific. The phosphoenolpyruvate carboxykinase activity assay kit (K359) was purchased from BioVision. PCK1(pS90) peptide (DVARIE-pS-KTIVIT), INSIG1/2(pS207/S151) peptide (WWTfDR-pS-RSGLGL) and PCK1(S90) covering peptide (WWTfDRSRSLGL) were synthesized by Selleck-Chem. CIP was obtained from New England BioLabs. Ni-NTA agarose was obtained from Qiagen.

DNA construction and mutagenesis

PCR-amplified human wild-type PCK1, PCK2, INSIG1 and INSIG2 and IGF1R were cloned into pcDNA3.1/hygro(+)-Flag, -HA, -His, or -Myc, pCDH-CMV-MCSEF1-Puro-SFB or pET32a vectors. SCAP was cloned into pcDNA3.1/hygro(+)-Myc. pECE-Myr-HA-AKT1(delta4-129), MSCV-XZ066-EGFR vIII and pBabe-puro-KRAS(G12V) were purchased from Addgene. EGFR vIII and KRAS(G12V) were cloned into pcDNA3.1/hygro(+)-Flag. Flag/His-double-tagged INSIG1 and INSIG2 were constructed in a pFastBacHTa expression vector (Invitrogen) as described²⁴. pT3-EF1a-c-Met was a gift from X. Chen (Addgene plasmid 31784). pCMV(CAT)T7-SB100 was a gift from Z. Izsvak (Addgene plasmid 34879). Flag-PCK1 (wild-type or S90A-mutant) and HA-myr-AKT were cloned into a pT3-EF1a vector. Myc-SCAP was a gift from Y. Chen at Shanghai Institute of Biological Sciences, Chinese Academy of Sciences.

pcDNA3.1 Flag-rPCK1, Flag-rPCK1(S90A), Flag-rPCK1(S90E), SFB-PCK1(S90A), pGEX-4T-1 PCK1(S90A), HA-AKT-DN (K179A, T308A, S473A), Flag-rPCK1(C288S), HA-rPCK1(C288S), PGEX-4T-1 PCK1(C228S), pET22b PCK1(S90A), pET22b PCK1(C288S), SFB-INSIG1(S207A), SFB-INSIG2(S151A), Flag-IGF1R(V922E), Flag-IGF1R(L1003R), Flag/His-double-tagged INSIG1(S207A) and INSIG1(S207E), Flag-INSIG1(S207D), Flag/His-double-tagged INSIG2(S151A) or INSIG2(S151E), Flag-INSIG2(S151D) and short hairpin RNA (shRNA)-resistant PCK1 constructs containing nonsense mutations of G948A, T951G, C954T and A957C were constructed using a QuikChange site-directed mutagenesis kit (Stratagene). pGIPZ shRNA was constructed via ligation of an oligonucleotide targeting human *PCK1* into an XhoI/MluI-digested pGIPZ vector. The following pGIPZ shRNA target sequences were used: control shRNA oligonucleotide, 5'-GCTTCTAACACCGGAGGTCTT-3'; *PCK1* shRNA oligonucleotide, 5'-TGTCGCTCAAACCTTCATCC-3'. *INSIG1* shRNA oligonucleotides, 5'-TAATGGTGTCTATCAGTATAC-3' and 5'-GGAACATAGGACGACAGTTA-3'; *INSIG2* shRNA oligonucleotides, 5'-CATCTAGGAGAACCTCATAAA-3' and 5'-CTTCAGCTGTGATTGGGTT-3'; *SCAP* shRNA oligonucleotides, 5'-CTCTTCAGCTATTACAACA-3' and 5'-AGGAAGAGGATGGTCTCCT-3'.

Cell lines and cell culture conditions

Hep3B, Huh7, H1993, CHL-1, SNU-398, SNU-475, HL7702, THLE-2 and 293T cells were from ATCC. The cells were maintained in complete medium containing Dulbecco's modified Eagle's medium (DMEM), 10% fetal bovine serum (FBS), 1,000 U ml⁻¹ penicillin and 100 µg ml⁻¹ streptomycin. Lipid-depleted medium contains DMEM supplemented with 5% LPDS and lovastatin (5 µM). Before IGF1 (100 ng ml⁻¹) treatment, the cells were serum-starved for 16 h. U0126 (20 µM), SU6656 (4 µM), SP600125 (25 µM), or MK-2206 (10 µM) was added 30 min before treatment with or without IGF1 (100 ng ml⁻¹) for 1 h. No cell lines used in this study were found in the database of commonly misidentified cell lines that is maintained by the International Cell Line Authentication Committee and NCBI Biosample. Cell lines were authenticated by short tandem repeat profiling and were routinely tested for mycoplasma contamination at The University of Texas MD Anderson Cancer Center. Cells were plated at a density of 4 × 10⁵ per 60-mm dish or 1 × 10⁵ per well of a 6-well plate 18 h before transfection. The transfection procedure was performed as previously described^{1,2}.

Immunoprecipitation and immunoblotting analysis

The extraction of proteins using a modified buffer from cultured cells was followed by immunoprecipitation and immunoblotting using corresponding antibodies as described previously^{1,2}.

GST pull-down assay

Equal amounts of His-tagged purified protein (200 ng per sample) were incubated with 100 ng of GST fusion proteins together with glutathione agarose beads in a modified binding buffer (50 mM Tris-HCl at pH 7.5, 1% Triton X-100, 150 mM NaCl, 1 mM DTT, 0.5 mM EDTA, 100 µM PMSF, 100 µM leupeptin, 1 µM aprotinin, 100 µM sodium orthovanadate, 100 µM sodium pyrophosphate, 1 mM sodium fluoride). The glutathione agarose beads were then washed four times with binding buffer and then subjected to immunoblotting analysis as previously described²⁵.

Purification of recombinant proteins

Wild-type GST-PCK1, GST-PCK1(S90A), GST-PCK1(C288S), His-PCK1, His-PCK1(S90A), His-PCK1(C228S) were expressed in bacteria and purified as described previously²⁶. Flag/His-double-tagged INSIG1 and INSIG2 were purified as previously described⁵.

[^3H]25-hydroxycholesterol binding assay

[^3H]25-hydroxycholesterol binding reactions were performed as previously described²⁴. In brief, each reaction was performed in a final volume

of 100 µl of buffer A (50 mM Tris-HCl at pH 7.5, 150 mM NaCl, 1 mM dithiothreitol, 0.1% Fos-choline 13, 0.005% sodium azide). It contained 1.2 µg (400 nM) of purified wild-type or mutants of Flag/His-double-tagged INSIG1 or INSIG2, 10–500 nM [³H]25-hydroxycholesterol and 25 mM phosphocholine chloride. After incubation for 4 h at room temperature, the mixture was passed through a column packed with 0.3 ml of Ni-NTA agarose beads (Qiagen). Each column was then washed 3 times with 10 ml of buffer B (50 mM Tris-HCl at pH 7.5, 150 mM NaCl, 1 mM dithiothreitol and 0.1% (w/v) Anapoe-C12E9). The protein-bound [³H]25-hydroxycholesterol was eluted with 250 mM imidazole and measured by scintillation counting.

Cell fractionation

The cell fraction assay was performed according to a previously reported method²⁷. In brief, 40 10-cm dishes of parental Huh7 cells and the indicated clones of cells with INSIG1(S207A)/INSIG2(S151A) double knock-in expression or PCK1(S90A) knock-in expression were stimulated with or without IGF1 (100 ng ml⁻¹) or incubated with or without lipid-depleted medium for 16 h. The cells were then placed on ice and washed twice with PBS and homogenization buffer (10 mM triethanolamine-acetic acid, pH 7.4, 0.25 M sucrose, 1 mM sodium EDTA, protease inhibitor cocktail (Roche)). The washed cells were collected in 0.8 ml homogenization buffer and homogenized by passing through a 25-gauge needle on a 1-ml syringe 13 times. After centrifugation at 2,000g for 15 min at 4 °C, the post-nuclear supernatant was collected and loaded on preformed iodixanol (Sigma-Aldrich) gradients. The discontinuous iodixanol gradients were prepared as 2.65 ml of 24%, 19.33%, 14.66% and 10%, which were made by diluting a 60% stock of iodixanol with cell suspension medium (0.85% (w/v) NaCl, 10 mM Tricine-NaOH, pH 7.4). After standing at room temperature for 2 h, the gradients were then centrifuged at 37,000 rpm in a SW40Ti rotor (Beckman Instruments) for 4 h. The post-nuclear supernatant was loaded on the top of the gradients and centrifuged at 37,000 rpm for another 2 h. Deceleration was performed without a brake. A total of 15 fractions (800 µl per fraction) were collected from the top to the bottom, and the bottom two fractions containing aggregated material were not analysed further. Aliquots of each fraction were used for further analysis by immunoblotting. ER fractions were isolated from the cells using an Endoplasmic Reticulum Isolation Kit (ER0100, Sigma-Aldrich). ER proteins were used in immunoblot analyses.

Conversion of radiolabelled glucose to triglycerides and fatty acids

The incorporation of the various radioactive substrates into triglycerides and fatty acids was measured as previously described²⁸. In brief, cells were washed twice with PBS and incubated in 1 ml labelling medium (2.5% fatty-acid-free bovine serum albumin, 1% (v/v) penicillin/streptomycin, 0.5 mM D-glucose, 0.5 mM sodium acetate, 2 mM sodium pyruvate, 2 µCi ml⁻¹ ¹⁴C-U-glucose or ¹⁴C-acetate) at 37 °C in a humidified incubator (5% CO₂) for 4.5 h before lipid extraction. All metabolic processes were stopped by washing cells twice with cold PBS and lysing them with the addition of modified Dole's extraction mixture (80 ml isopropanol, 20 ml hexane, 2 ml 0.5 M H₂SO₄). Triglycerides were extracted with hexane and washed, and the solvent was evaporated. The incorporation of ¹⁴C-glucose into fatty acids and triglycerides was determined by evaporating the solvent from neutral lipids, adding 1 ml of KOH-ethanol (20 ml 95% ethanol, 1 ml water, 1 ml saturated KOH) and heating samples to 80 °C for 1 h. Sulfuric acid was added to the mixture to ensure complete saponification. Addition of hexane allowed for hydrophobic separation. The hydrophilic portion was evaporated and counted using liquid scintillation. Incorporation data were normalized according to cell number.

In vitro kinase assay

In vitro kinase assays were performed as previously reported²⁹. In brief, for the AKT in vitro kinase assay, purified active GST-AKT

(A16-10G, SignalChem) (500 ng) was incubated with bacterially purified His-PCK1 (200 ng) in 25 µl kinase buffer (50 mM Tris-HCl at pH 7.5, 100 mM KCl, 50 mM MgCl₂, 1 mM Na₃VO₄, 1 mM DTT, 5% glycerol, 0.5 mM ATP and 10 µCi [³²P]ATP) at 25 °C for 1 h. The reaction was terminated by adding SDS-PAGE loading buffer and heated at 100 °C for 5 min. The reaction mixture was then subjected to an SDS-PAGE analysis. For PCK1 in vitro kinase assay, bacterially purified wild-type His-PCK1 or His-PCK1(S90A) on the Ni-NTA agarose beads was incubated with or without GST-AKT1 in the presence of ATP for 1 h. After the in vitro AKT kinase assay, the Ni-NTA agarose beads were washed in PBS five times and incubated with or without SFB-tagged wild-type or mutant INSIG1/2 using 50 µl kinase buffer in the presence of 0.5 mM GTP and 10 µCi [³²P]ATP or [³²P]GTP at 25 °C for 1 h. Autoradiography was performed.

Mass spectrometry analyses

For identification of interacting proteins, a protein band visualized via Coomassie blue staining was excised from an SDS-PAGE gel and digested in gel in 50 mM ammonium bicarbonate buffer containing RapiGest (Waters Corporation) overnight at 37 °C with 200 ng of modified sequencing-grade trypsin (Promega). The digested protein samples were analysed using high-sensitivity LC-MS/MS with an Orbitrap Elite mass spectrometer (Thermo Fisher Scientific). Proteins were identified by searching the fragment spectra against the UniProt protein database (EMBL-EBI) using the Mascot search engine (v.2.3; Matrix Science) with the Proteome Discoverer software program (v.1.4; Thermo Fisher Scientific). For detection of phosphorylation sites, in vitro phosphorylation of PCK1 by AKT and INSIG1/2 by PCK1 were performed according to the in vitro kinase assay protocol described above. Then the protein samples were digested using trypsin or chymotrypsin and analysed using LC-MS/MS with the Orbitrap Elite mass spectrometer as described previously³⁰.

Determining the K_m of PCK1

The K_m of PCK1 was determined using a GTPase assay kit (MAK113, Sigma-Aldrich) according to the manufacturer's instructions. In brief, purified recombinant wild-type PCK1 (10 ng) was incubated in 100 µl of reaction buffer (40 mM Tris-HCl at pH 7.5, 80 mM NaCl, 10 mM synthetic INSIG1 peptide substrate (LWWTFD RSRSLGLG), 8 mM magnesium acetate, 1 mM DTT) with different concentrations of GTP at 37 °C in 96-well plates. The plates were read by multi-detection microplate readers (BMG Labtech) at 620 nm in kinetic mode for 5 min. K_m and V_{max} were calculated from a plot of 1/V versus 1/[substrate] according to the Lineweaver-Burke plot model.

Determination of oxaloacetate binding affinity and detection of PCK1 activity

Ni-NTA beads and purified recombinant wild-type His-PCK1, His-PCK1(S90E) or AKT-phosphorylated His-PCK1 were incubated in 200 µl binding buffer (50 mM Tris-HCl at pH 7.5, 100 mM KCl, 50 mM MgCl₂, 50 mM MnCl₂, 1 mM Na₃VO₄) containing 200 µM oxaloacetate (O7753, Sigma-Aldrich) at 30 °C for 30 min. The oxaloacetate associated with His-PCK1 on the beads was then centrifuged at 12,000 rpm for 10 min. The oxaloacetate remaining in the supernatant was quantified using an oxaloacetate assay kit (ab83428, Abcam). The oxaloacetate binding affinity was calculated according to the percentages of the remaining oxaloacetate in the supernatants. The activity of PCK1 was determined using a phosphoenolpyruvate carboxykinase activity assay kit (K359, BioVision).

CRISPR-Cas9-mediated genome editing

Genomic mutations were introduced into cells using the CRISPR-Cas9 system, as described previously³⁰. Single-guide RNAs (sgRNAs) were designed to target the genomic area adjacent to mutation sites in PCK1(S90A), INSIG1(S207A) and INSIG2(S151A) using the CRISPR design tool (<http://crispr.mit.edu/>). The annealed guide RNA oligonucleotides were inserted into a PX458 vector (Addgene) digested with

Article

the BbsI restriction enzyme³¹. Cells were seeded at 60% confluence, followed by co-transfection of sgRNAs (0.5 µg) and single-stranded donor oligonucleotide (10 pmol) as a template to introduce mutations. Twenty-four hours after transfection, cells were trypsinized, diluted for single cells and seeded into 96-well plates. Genomic DNA was extracted from GFP-positive cells, followed by sequencing of the PCR products spanning the mutation sites. sgRNA targeting sequence for INSIG1(S207A): 5'-ACATTTGATCGTTCCAGAAG-3'; single-stranded donor oligonucleotide (ssODN) sequence for INSIG1(S207A): 5'-AAATTGGATTTTGCCAATAATGTCCAGCTGTCCTTGACTTTAGCAGCCCTATCTTTGGGCCCTTTGGTGGACATTTGATCGcgcCAGgAgcGGCCTTGGGCTGGGGATCACCATAGCTTTTCTAGCTACGCTGATCACGCAGTTTCTCGTGATAATGGTGTCTATCA-3'; ssODN sequence for INSIG1(S207D): 5'-AAATTGGATTTTGCCAATAATGTCCAGCTGTCCTTGACTTTAGCAGCCCTATCTTTGGGCCCTTTGGTGGACATTTGATCGcgaCAGgAgcGGCCTTGGGCTGGGGATCACCATAGCTTTTCTAGCTACGCTGATCACGCAGTTTCTCGTGATAATGGTGTCTATCA-3'; sgRNA targeting sequence for INSIG2(S151A): 5'-ACTTTTGATAGATCTAGAAG-3'; ssODN sequence for INSIG2(S151A): 5'-AAAGTGGATTTGATAACAACATACAGTTGTCTCTCACACTGGCTGCACTATCCATTGGACTGTGGTGGACTTTTGATAGggCTAGgAgcGGTTTGGCCCTGGAGTAGGAATTGCCTTCTTGGCAACTGTGGTCACTCAACTGCTAGTATAATGGTGTTCACCA-3'. ssODN sequence for INSIG2(S151D): 5'-AAAGTGGATTTGATAACAACATACAGTTGTCTCTCACACTGGCTGCACTATCCATTGGACTGTGGTGGACTTTTGATAGggaTAGgAgcGGTTTGGCCCTGGAGTAGGAATTGCCTTCTTGGCAACTGTGGTCACTCAACTGCTAGTATAATGGTGTTCACCA-3'. sgRNA targeting sequence for PCK1(S90A): 5'-GGCCAGGATCGAAAGCAAGA-3'; ssODN sequence for PCK1(S90A): 5'-CCGTGGTGCTTGGCTGA AAGGAAGCCTGTGA TTTTTCAGCTGGTGGCTCTCACTGACCC CAGGGATGTGGCCAGGATaGAggcCAAaACGGTTATCGTCACCCAAGA GCAAAGAGACACAGTGCCCATCCCAAAACAGGCCTAGCCAGCTCG GTCGCTGGATGTCAGAGG-3'. The lower-case letters in the ssODN sequences indicate the mutated nucleotides that will replace the endogenous nucleotides in the genomic DNA of parental cells using the CRISPR-Cas9 system. Genotyping was performed by sequencing PCR products amplified from the following primers: INSIG1 forward: 5'-AGAATGGGGCTATCGATGACTTC-3'; INSIG1 reverse: 5'-TGTAGTGGGGATATGCAGAACG-3'; INSIG2 forward: 5'-TC AAGTTCCTGTACGATTCTCAAGT-3'; INSIG2 reverse: 5'-AGCAA ACAAGCACCAAAAATTG-3'; PCK1 forward: 5'-AAGGCCTTCG GGTAGTTTCAG-3'; PCK1 reverse: 5'-AGCCCCCTGGGTTAGAAGAG-3'.

SRE luciferase assay

SRE luciferase activity in cell lysates was measured with the luciferase assay system as previously described³². In brief, Huh7 cells in 24-well plates were transfected with 0.1 µg SRE-driven luciferase reporter and 0.075 µg β-galactosidase. At 24 h after the transfection, the cells were subjected to different culture media. Cell lysates were measured for the activity of luciferase and β-galactosidase on the basis of the manufacturer's instruction (Promega).

Immunofluorescence analysis

Immunofluorescence analysis was performed as previously reported³³. Cultured cells were fixed by 4% paraformaldehyde (PFA), treated with 0.1% Triton X-100 for 5 min and blocked in 3% BSA for 1 h. The cells were then incubated with primary antibodies at a dilution of 1:100. For tissue staining, tumour masses from mice were perfused with 0.1 M PBS (pH 7.4), embedded into optimal cutting temperature compound and frozen for cryostat section. Cryostat sections were fixed with 4% PFA for 15 min at room temperature. After PBS washing, cryostat sections were incubated in the blocking solution (PBS containing 3% donkey serum, 1% BSA, 0.3% Triton X-100 at pH 7.4) for 30 min at room temperature. In antibody reaction buffer (PBS plus 1% BSA, 0.3% Triton X-100 at pH 7.4), samples were stained with primary antibodies against SREBP1 antibody overnight at 4 °C. After incubation with fluorescent-dye-conjugated

secondary antibodies and DAPI, immunofluorescent microscopic images of the cells were obtained and viewed using an IX81 confocal microscope (Olympus America). Colocalization of proteins was quantified by calculating Pearson's correlation coefficient using the Coloc 2 plugin in Image J (National Institutes of Health).

Cell viability analysis

Cells (2×10^5) were plated in DMEM with 10% FBS (complete medium). After treatment with lipid-depleted medium for the indicated time, the viable cells were stained with trypan blue (0.5%) and counted using a Cell Viability Analyzer (Beckman Coulter).

Quantitative PCR

Total RNA was extracted from cells and tissue samples using TRIzol reagent according to the manufacturer's instructions (Invitrogen). Equal amounts of RNA samples were used for cDNA synthesis with a TaqMan Reverse Transcription Reagents kit (Applied Biosystems). Quantitative PCR analysis was carried out using a 7500 Real-Time PCR system (Applied Biosystems) with a SYBR Premix Ex Taq kit (Takara Bio). The following primers were used for quantitative PCR: *FASN*, 5'-CACAGGGACAACCTGGAGTT-3' and 5'-ACTCCACAGGTGGGAACAAG-3'; *SCD*, 5'-CGACGTGGCTT TTTCTTCTC-3' and 5'-CCTTCTCTTTGACAGCTGGG-3'; *ACACA*, 5'-AGTGGGTCACCCATTGTT-3' and 5'-TTCTAACAGGAGCTGGAGCC-3'; *GPAM*, 5'-TTGTGGCTTGCTGCTCCTCTA-3' and 5'-AATCACGAG CCAGGACTTCCTC-3'; *HMGR*, 5'-TCTGGCAGTCAGTGGGAACATT-3' and 5'-CCTCGTCCCTTCGATCCAATTT-3'; *HMGCS1*, 5'-GATGTG GGAATTGTTGCCCTT-3' and 5'-ATTGTCTCTGTTCCAACCTCCAG-3'; *LDLR*, 5'-AACGGTCATTACCCAGGTC-3' and 5'-GGCTGAAGA ATAGGAGTTGCC-3'; *FDFT1*, 5'-CGATAGCTGTGTGCAAAGTAAC-3' and 5'-CCATCTGCTGAGTGCTTTCTG-3'; *GAPDH*, 5'-AGCCACATCGC TCAGACAC-3' and 5'-GCCCCAATACGACCAATCC-3'.

TUNEL assay

Mouse tumour tissues were sectioned at 5-µm thickness. Apoptotic cells were counted using the DeadEnd Colorimetric TUNEL System (Promega) according to the manufacturer's instructions.

IHC analysis and histological evaluation of human HCC specimens

Human HCC tissue collection and study approval were described previously³⁴. Human HCC and adjacent matched non-tumour tissue samples (EHBH cohort) were obtained from Eastern Hepatobiliary Surgery Hospital in Shanghai, China. The use of human HCC samples and the relevant database was approved by the Eastern Hepatobiliary Surgery Hospital Research Ethics Committee and complied with all relevant ethical regulations. All tissue samples were collected in compliance with the informed consent policy. Sections of paraffin-embedded human HCC samples were stained with antibodies against AKT (pS473), PCK1 (pS90), INSIG1 (pS207)/INSIG2 (pS151), SREBP1 or non-specific IgG as a negative control. The staining of the tissue sections was quantitatively scored according to the percentage of positive cells and the staining intensity as described previously³⁰. The following proportion scores were assigned to the sections: 0 if 0% of the tumour cells exhibited positive staining, 1 for 0–1%, 2 for 2–10%, 3 for 11–30%, 4 for 31–70% and 5 for 71–100%. In addition, the staining intensity was rated on a scale of 0–3: 0, negative; 1, weak; 2, moderate; and 3, strong. The proportion and intensity scores were then combined to obtain a total score (range, 1–8) as described previously³⁰. Scores were compared with overall survival duration, defined as the time from date of diagnosis to that of death or last known follow-up examination. All patients had received standard therapies after surgery.

Mouse studies

One million Huh7 cells with or without gene editing in PCK1 and INSIG1/2 were collected in 20 µl DMEM with 33% matrigel and intrahepatically

or subcutaneously injected into 6-week-old male BALB/c athymic nude mice. The injections were performed as described previously³⁵. Seven mice per group in each experiment were used. Mice were euthanized 28 days after injection. The liver of each mouse was dissected and then fixed in 4% formaldehyde and embedded in paraffin. The tumour volume was calculated using the formula: $V = 1/2a^2b$ (V , volume; a , shortest diameter; b , longest diameter). The mice were treated in accordance with relevant institutional and national guidelines and regulations.

The transgene HCC mouse model was established by overexpression of activated AKT combined with c-Met using hydrodynamic transfection²². Wild-type FVB/N mice were subjected to hydrodynamic injection as previously described³⁶. In brief, the plasmids pT3-EF1 α -Flag-PCK1 (wild-type or S90A-mutant) (20 μ g), pT3-EF1 α -HA-myr-AKT (20 μ g) and pT3-EF1 α -V5-c-Met (20 μ g) together with pCMV/sleeping beauty transposase (SB) (2.4 μ g), in a ratio of 12.5:12.5:12.5:1.5, were diluted in 2 ml saline (0.9% NaCl), filtered through a 0.22- μ m filter and injected into the lateral tail vein of the male mice in 6 s. After 14 weeks, the mice were euthanized and liver tumours were extracted.

The use of the mice was approved by the Institutional Review Board at MD Anderson Cancer Center and the Institutional Animal Care and Use Committee (IACUC) of Zhejiang University and Qingdao Cancer Institute, China. The maximum tumour diameter permitted by the committees is 1.5 cm. Mice arriving in the animal facility were randomly put into cages with five mice each. No statistical methods were used to predetermine sample size.

Statistics and reproducibility

All statistical data are presented as means \pm s.d. All experiments were repeated at least twice independently with similar results. The mean values obtained in the control and experimental groups were analysed for significant differences. Pairwise comparisons were performed using a two-tailed t test. P values of less than 0.05 were considered significant. Unless stated otherwise, the experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All source data for immunoblotting are shown in Supplementary Fig. 1. All the other source data for Figs. 2–4 and Extended Data Figs. 1–10, containing the raw data for all experiments, are provided with the paper. All images and data were created and analysed by the authors and will

be available from the lead corresponding author (Z.L.) on reasonable request.

24. Radhakrishnan, A., Sun, L. P., Kwon, H. J., Brown, M. S. & Goldstein, J. L. Direct binding of cholesterol to the purified membrane region of SCAP: mechanism for a sterol-sensing domain. *Mol. Cell* **15**, 259–268 (2004).
25. Xu, D. Q. et al. PAQR3 controls autophagy by integrating AMPK signaling to enhance ATG14L-associated PI3K activity. *EMBO J.* **35**, 496–514 (2016).
26. Xia, Y. et al. c-Jun downregulation by HDAC3-dependent transcriptional repression promotes osmotic stress-induced cell apoptosis. *Mol. Cell* **25**, 219–232 (2007).
27. Hammond, C. & Helenius, A. Quality control in the secretory pathway: retention of a misfolded viral membrane glycoprotein involves cycling between the ER, intermediate compartment, and Golgi apparatus. *J. Cell Biol.* **126**, 41–52 (1994).
28. Danai, L. V. et al. Map4k4 suppresses Srebp-1 and adipocyte lipogenesis independent of JNK signaling. *J. Lipid Res.* **54**, 2697–2707 (2013).
29. Lee, J. H. et al. EGFR-phosphorylated platelet isoform of phosphofructokinase 1 promotes PI3K activation. *Mol. Cell* **70**, 197–210 (2018).
30. Li, X. et al. Mitochondria-translocated PGK1 functions as a protein kinase to coordinate glycolysis and the TCA cycle in tumorigenesis. *Mol. Cell* **61**, 705–719 (2016).
31. Sander, J. D. & Joung, J. K. CRISPR–Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.* **32**, 347–355 (2014).
32. Xu, D. et al. PAQR3 modulates cholesterol homeostasis by anchoring Scap/SREBP complex to the Golgi apparatus. *Nat. Commun.* **6**, 8100 (2015).
33. Jin, T. et al. Identification of the topology and functional domains of PAQR10. *Biochem. J.* **443**, 643–653 (2012).
34. Xu, D. et al. The protein kinase activity of fructokinase A specifies the antioxidant responses of tumor cells by phosphorylating p62. *Sci. Adv.* **5**, eaav4570 (2019).
35. Li, X. et al. A splicing switch from ketohexokinase-C to ketohexokinase-A drives hepatocellular carcinoma formation. *Nat. Cell Biol.* **18**, 561–571 (2016).
36. Calvisi, D. F. et al. Increased lipogenesis, induced by AKT-mTORC1-RPS6 signaling, promotes development of human hepatocellular carcinoma. *Gastroenterology* **140**, 1071–1083 (2011).

Acknowledgements We thank L. Li for technical assistance. The mass spectrometry was supported in part by the Clinical and Translational Proteomics Service Center at the University of Texas Health Science Center. This work was supported by Zhejiang University Research Fund (188020*194221901/029) (Z.L.); the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang (2019R01001) (Z.L.); MOHW109-TDU-B-212-010001 (M.-C.H.); the Drug Development Center, China Medical University from the Ministry of Education in Taiwan (M.-C.H.); and The Odyssey Fellowship from The University of Texas MD Anderson Cancer Center (D.X.). Z.L. is Kuancheng Wang Distinguished Chair.

Author contributions Z.L. and D. Xu conceived and designed the study and wrote the manuscript; Z.L. and M.-C.H. acquired the funding support and supervised the study; D. Xu and Z.W. performed most experiments; Y.X., F.S., X.L. and X.Q. provided support for generating different knock-in mutation cell lines and protein purification; W.X., Y.W. and G.L. provided support for IHC staining; F.S., J.-H.L. and L.D. were involved in the studies with mice; M.-C.H., D. Xing and T.L. reviewed and edited the manuscript; H.W. provided the HCC samples and technical support; Y.Z., J.-s.L., D. Xing and T.L. performed statistical analysis and interpretation of the data.

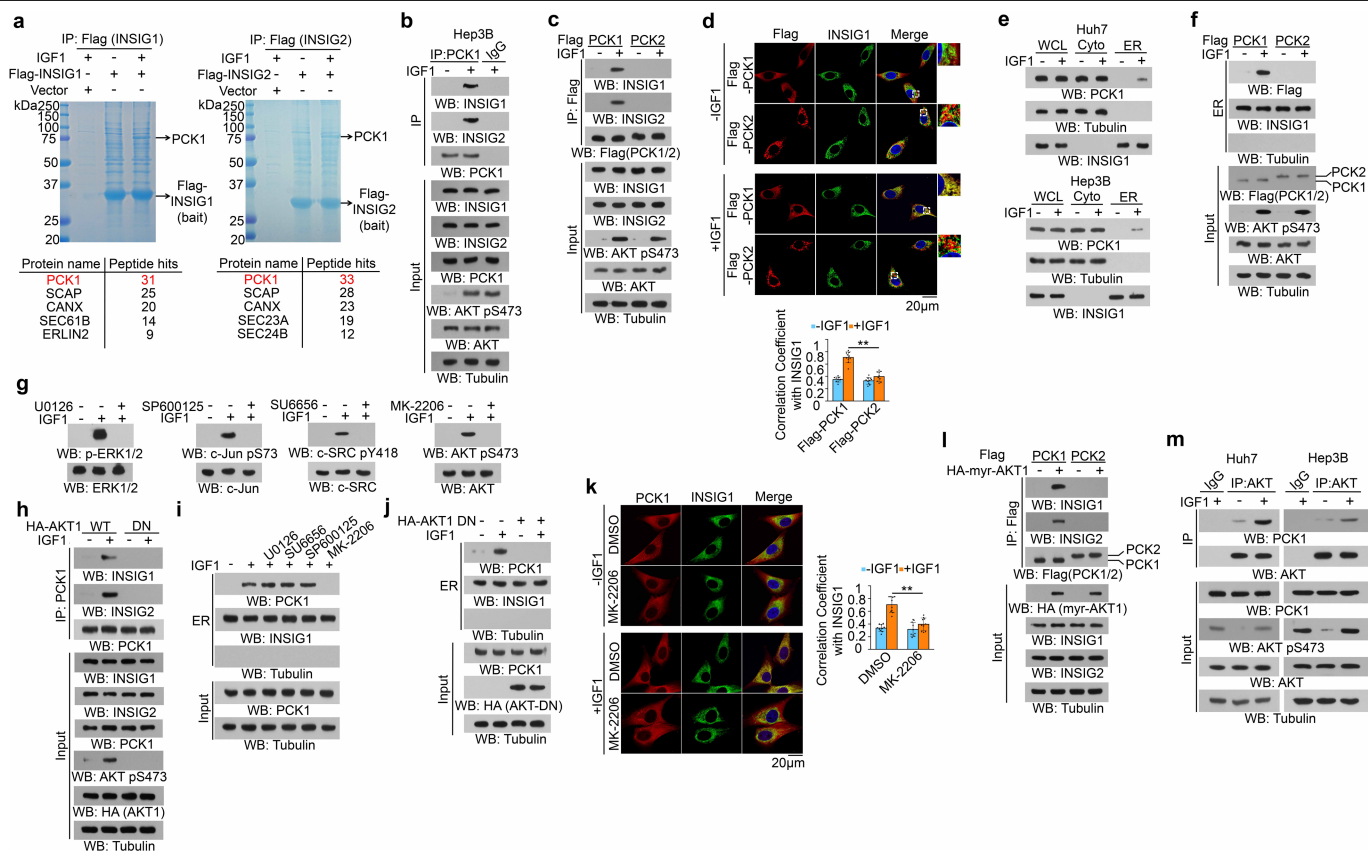
Competing interests Z.L. owns shares in Signalway Biotechnology (Pearland, TX), which supplied rabbit antibodies that recognize PCK1(pS90), INSG1(pS207) and INSG2(pS151). The interest of Z.L. in this company had no bearing on its being chosen to supply these reagents.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2183-2>.

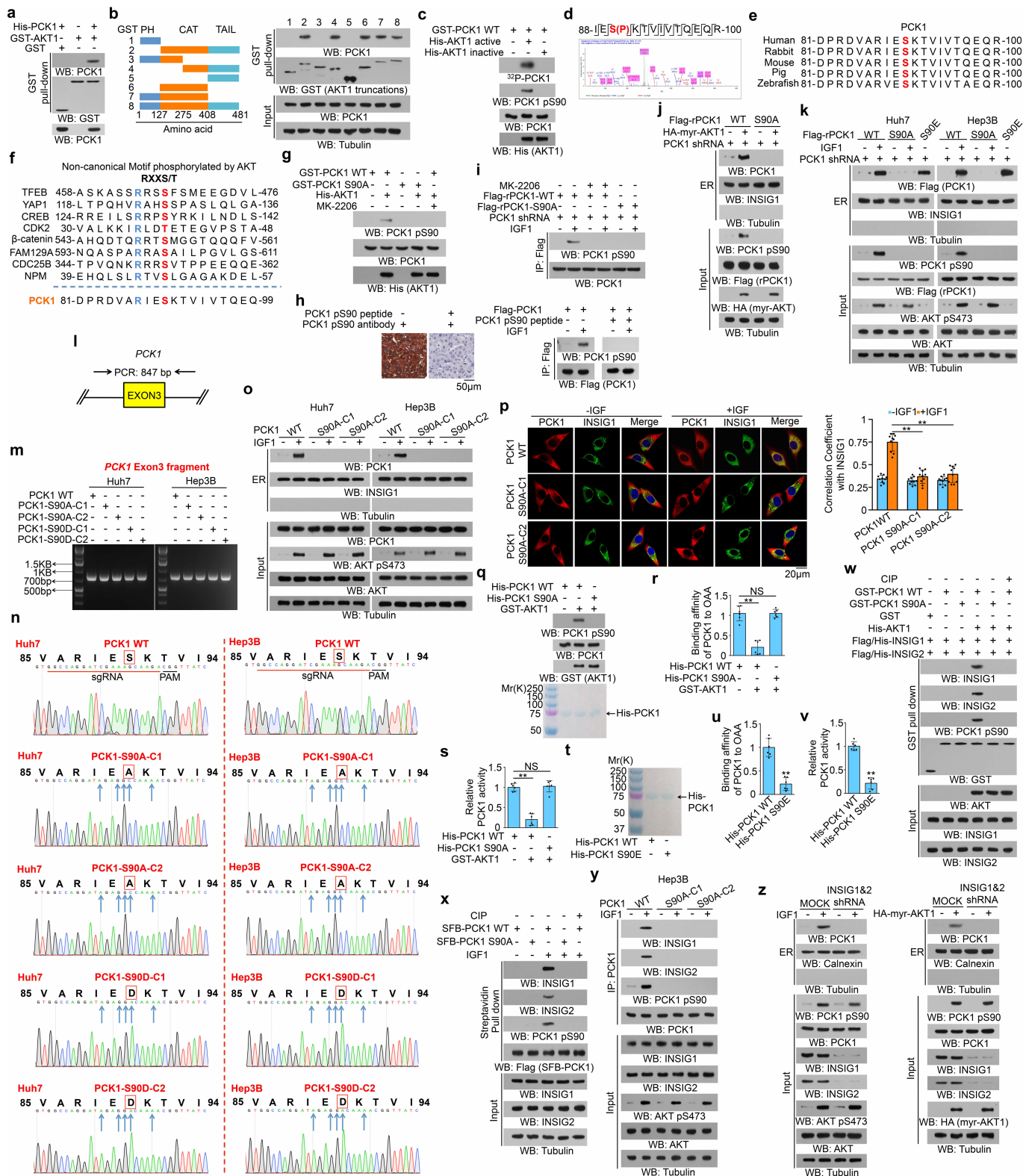
Correspondence and requests for materials should be addressed to D.X., M.-C.H. or Z.L.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



Extended Data Fig. 1 | IGF1-induced AKT activation induces the translocation of PCK1 to the ER and the binding of PCK1 to INSIG1/2. **a**, Huh7 cells with or without expression of Flag-INSIG1 (left) or Flag-INSIG2 (right) were treated with or without IGF1 (100 ng ml⁻¹) for 1 h. An immunoprecipitation assay was performed using anti-Flag antibody, and immunoprecipitates of Flag-INSIG1 or Flag-INSIG2 were eluted with Flag peptide, separated using SDS-PAGE and stained with Coomassie Brilliant Blue. Selected peptide hits of proteins associated with Flag-INSIG1 or Flag-INSIG2, identified through mass spectrometry, are shown. **b**, Hep3B cells were treated with or without IGF1 (100 ng ml⁻¹) for 1 h. **c**, Huh7 cells expressing Flag-PCK1 or Flag-PCK2 were stimulated with or without IGF1 (100 ng ml⁻¹) for 1 h. **d**, Huh7 cells expressing Flag-PCK1 or Flag-PCK2 were stimulated with or without IGF1 (100 ng ml⁻¹) for 1 h. Immunofluorescence analyses were performed with the indicated antibodies (top). The colocalization coefficients between the indicated proteins in the presence or absence of IGF1 are shown (bottom). At least $n = 50$ cells from each independent experiment were analysed and representative data are shown. Data are mean \pm s.d. $^{**}P < 0.001$ (two-tailed t -test). The regions in white boxes are shown at higher magnification on the right. **e**, Whole cell lysate (WCL), cytosolic and ER fractions were prepared from Huh7 and Hep3B cells stimulated with or without IGF1 (100 ng ml⁻¹) for 1 h. Cellular fractions from equal numbers of cells were analysed using immunoblotting with the indicated antibodies. **f**, Huh7 cells expressing Flag-PCK1 or Flag-PCK2 were stimulated with or without IGF1 (100 ng ml⁻¹) for 1 h. ER fractions and total lysates were prepared for immunoblotting analyses with the

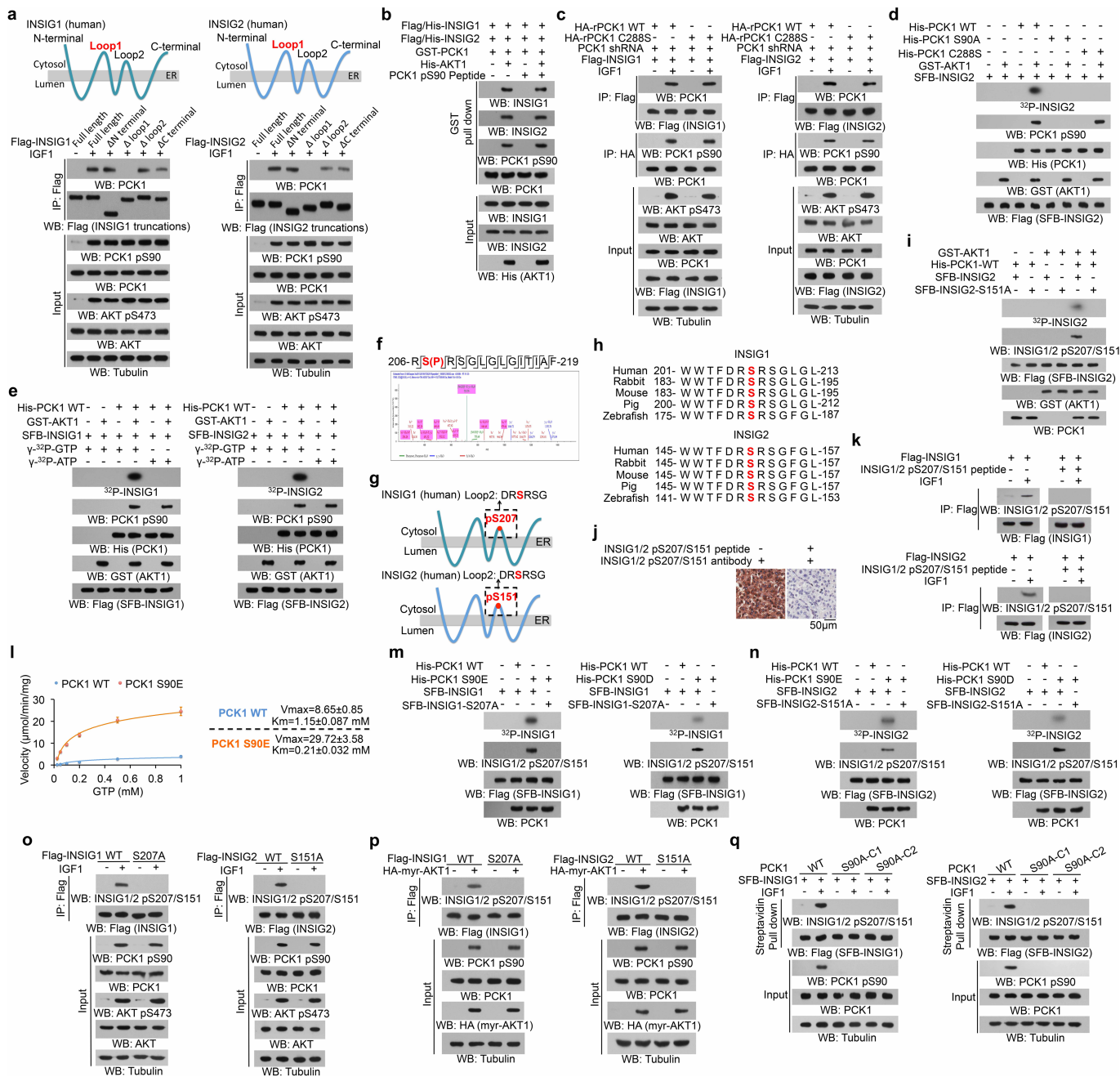
indicated antibodies. **g**, Total lysates were prepared from Huh7 cells pretreated with or without U0126 (20 μ M), SP600125 (25 μ M), SU6656 (4 μ M) or MK-2206 (10 μ M) for 30 min before treatment with or without IGF1 (100 ng ml⁻¹) for 1 h. **h**, Huh7 cells expressing wild-type HA-AKT1 or HA-AKT1-DN (K179A, T308A, S473A) were treated with or without IGF1 (100 ng ml⁻¹) for 1 h. ER fractions and total lysates were isolated for immunoblotting analyses with the indicated antibodies. **i**, Huh7 cells were pretreated with or without U0126 (20 μ M), SU6656 (4 μ M), SP600125 (25 μ M) or MK-2206 (10 μ M) for 30 min before treatment with or without IGF1 (100 ng ml⁻¹) for 1 h. ER fractions and total lysates were isolated for immunoblotting analyses with the indicated antibodies. **j**, Huh7 cells were stably transfected with a control vector or a vector expressing HA-AKT1-DN (K179A, T308A, S473A). The cells were treated with or without IGF1 (100 ng ml⁻¹) for 1 h. ER fractions and total lysates were isolated for immunoblotting analyses with the indicated antibodies. **k**, Huh7 cells were pretreated with or without MK-2206 (10 μ M) for 30 min and treated with or without IGF1 (100 ng ml⁻¹) for 1 h. Immunofluorescence analyses were performed with the indicated antibodies (left). The colocalization coefficients between the indicated proteins are shown (right). At least $n = 50$ cells from each independent experiment were analysed and representative data are shown. Data are mean \pm s.d. $^{**}P < 0.001$ (two-tailed t -test). **l**, Huh7 cells expressing Flag-PCK1 or Flag-PCK2 were transfected with or without HA-myr-AKT1. **m**, Huh7 and Hep3B cells were treated with or without IGF1 (100 ng ml⁻¹) for 1 h. In **b**, **c**, **e**, **j**, **l**, **m**, immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies, and the experiments were repeated three times independently with similar results.



Extended Data Fig. 2 | See next page for caption.

Extended Data Fig. 2 | AKT-mediated phosphorylation of PCK1 Ser90 is necessary and sufficient for the translocation of PCK1 to the ER and the binding of PCK1 to INSIG1/2, and inhibits the canonical role of PCK1 in gluconeogenesis. **a**, A GST pull-down assay was performed by mixing purified His-PCK1 with purified GST or GST-AKT1. Immunoblotting analyses were performed as indicated. **b**, Left, schematic of AKT1 full-length and deletion mutants. Right, 293T cells were transfected with the indicated constructs, and GST pull-down assays and immunoblotting analyses with the indicated antibodies were performed. **c**, In vitro kinase assays were performed by mixing purified GST-PCK1 with purified active or inactive His-AKT1 in the presence of [γ - 32 P]ATP. Autoradiography and immunoblotting analyses were performed as indicated. **d**, In vitro kinase assays were performed by mixing purified His-PCK1 with or without purified GST-AKT1 in the presence of ATP. Mass spectrometric analysis of a tryptic fragment at m/z 805.91107 Da (-0.06 mmu/ -0.08 ppm), which was matched with the +2 charged peptide 88-IESKTVIVTQEQR-100, suggested that PCK1 Ser90 was phosphorylated. The Mascot score was 31, and the expectation value was 0.25. **e**, Alignment of protein sequences spanning PCK1 Ser90 from different species. **f**, Alignment of PCK1 Ser90 to the non-canonical AKT-phosphorylated substrate motif (RXXS/T). The reported AKT substrates (TFEB, YAP1, CREB, CDK2, β -catenin, FAM129A, CDC25B and NPM) are shown. Red, phosphoacceptor residue; blue, basic residue (R). **g**, In vitro kinase assays were performed by mixing purified GST-PCK1 or GST-PCK1(S90A) and active His-AKT1 in the presence or absence of MK-2206 (10 μ M) for 1 h. Immunoblotting analyses were performed as indicated. **h**, Left, IHC analyses of human HCC samples were performed with the indicated antibodies in the presence or absence of a blocking peptide for PCK1(pS90). Right, Huh7 cells expressing Flag-PCK1 were treated with or without IGF1 (100 ng ml $^{-1}$) for 1 h. Immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies in the presence or absence of a blocking peptide for PCK1(pS90). **i**, Huh7 cells expressing wild-type Flag-rPCK1 or Flag-rPCK1(S90A) were pretreated with or without MK-2206 (10 μ M) for 30 min. The cells were treated with or without IGF1 (100 ng ml $^{-1}$) for 1 h. Immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies. **j**, PCK1-depleted Huh7 cells were reconstituted with the indicated shRNA-resistant PCK1 proteins. After transfection with HA-myr-AKT1, ER fractions and total cell lysate were prepared for immunoblotting analyses as indicated. **k**, Huh7 and Hep3B cells expressing the indicated Flag-rPCK1 proteins were treated with or without IGF1 (100 ng ml $^{-1}$) for 1 h. ER fractions and total cell lysate were prepared for immunoblotting analyses as indicated. **l**, **m**, Genomic DNA was extracted from two individual clones of parental Huh7 or Hep3B cells with knock-in expression of PCK1(S90A). PCR products were amplified from the

indicated DNA fragment (**l**) and separated on an agarose gel (**m**). **n**, Sequencing of parental Huh7 and Hep3B cells and two individual clones of parental cells with knock-in expression of PCK1(S90A). The red line indicates the sgRNA-targeting sequence. The black line indicates the protospacer adjacent motif (PAM). Blue arrows indicate mutated nucleotides. A mutated amino acid and its wild-type counterpart are indicated by the solid red box. **o**, **p**, Parental Huh7 and Hep3B cells and the indicated clones of Huh7 and Hep3B cells with knock-in expression of PCK1(S90A) were stimulated with or without IGF1 (100 ng ml $^{-1}$) for 1 h. ER fractions and total cell lysates were prepared and immunoblotting analyses were performed as indicated (**o**). Immunofluorescence staining of Huh7 cells was performed with the indicated antibodies (**p**, left). The colocalization coefficients between the indicated proteins are shown (**p**, right). At least $n = 50$ cells from each independent experiment were analysed and representative data are shown. Data are mean \pm s.d. $^{**}P < 0.001$ (two-tailed t -test). **q-s**, Bacterially purified wild-type His-PCK1 or His-PCK1(S90A) on Ni-NTA agarose beads were incubated with or without purified active GST-AKT1 in the presence of ATP for an in vitro AKT kinase assay. Immunoblotting analyses were performed as indicated (**q**). After washing wild-type His-PCK1 or His-PCK1(S90A)-conjugated beads with PBS five times, the binding affinity of PCK1 to oxaloacetate (OAA) (**r**) and the relative PCK1 activity (**s**) were measured. Data are mean \pm s.d. ($n = 6$ biological replicates). $^{**}P < 0.001$ (two-tailed t -test); NS, not significant ($P = 0.923$ (**r**); $P = 0.728$ (**s**)). **t-v**, Wild-type His-PCK1 and His-PCK1(S90E) were purified from bacteria and Coomassie Brilliant Blue staining analyses were performed (**t**). The binding affinity of the His-PCK1 proteins to oxaloacetate (**u**) and the relative PCK1 activity (**v**) were measured. Data are mean \pm s.d. ($n = 6$ biological replicates). $^{**}P < 0.001$ (two-tailed t -test). **w**, Bacterially purified wild-type GST-PCK1 or GST-PCK1(S90A) on glutathione agarose beads were incubated with or without active His-AKT1 in the presence of ATP for an in vitro kinase assay. The GST-tagged proteins were then incubated with or without CIP (10 U) for 30 min at 37 $^{\circ}$ C, followed by incubation with Flag/His-tagged INSIG1 or INSIG2 purified from Huh7 cells for a pull-down assay. **x**, SFB-tagged wild-type PCK1 or PCK1(S90A) were pulled down from Huh7 cells treated with or without IGF1 (100 ng ml $^{-1}$) for 1 h. These proteins were incubated with or without CIP (10 U) for 30 min at 37 $^{\circ}$ C. Immunoblotting analyses were performed as indicated. **y**, Parental Hep3B cells and the indicated clones of cells with knock-in expression of PCK1(S90A) were stimulated with or without IGF1 for 1 h. Immunoprecipitation and immunoblotting analyses were performed as indicated. **z**, Huh7 cells expressing *INSIG1* shRNA and *INSIG2* shRNA were treated with or without IGF1 (100 ng ml $^{-1}$) for 1 h (left) or transfected with HA-myr-AKT1 (right). ER fractions and total cell lysate were prepared for immunoblotting analyses as indicated. All experiments were repeated three times independently with similar results.



Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | INSIG1/2 is required for the translocation of PCK1 to the ER and PCK1 functions as a protein kinase to phosphorylate INSIG1 Ser207 and INSIG2 Ser151. **a**, Top, the topological structures of INSIG1 and INSIG2, which have 69% amino acid sequence identity and contain six transmembrane-spanning regions. Bottom, different INSIG1 (left) or INSIG2 (right) truncation mutants were expressed in 293T cells. These cells were treated with or without IGF1 (100 ng ml⁻¹) for 1 h. Immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies. **b**, Bacterially purified GST–PCK1 was incubated with or without active His–AKT1 in the presence of ATP for 1 h. After the in vitro kinase assay, GST–PCK1-conjugated beads were washed with PBS five times and then incubated with purified Flag/His–INSIG1 or Flag/His–INSIG2 in the absence or presence of a PCK1(pS90) peptide for 2 h. Immunoblotting analyses were performed with the indicated antibodies. **c**, Endogenous PCK1-depleted Huh7 cells with reconstituted expression of shRNA-resistant wild-type rPCK1 or rPCK1(C228S) were transfected with Flag–INSIG1 (left) or Flag–INSIG2 (right), respectively. After being treated with or without IGF1 (100 ng ml⁻¹) for 1 h, the cells were collected for immunoprecipitation and immunoblotting analyses as indicated. **d**, Bacterially purified wild-type His–PCK1, His–PCK1(S90A) or His–PCK1(C288S) on Ni-NTA agarose beads were incubated with or without active GST–AKT1 in the presence of ATP for an in vitro AKT kinase assay. The beads were washed with PBS five times and incubated with SFB–INSIG2 in the presence of [γ -³²P]GTP. Autoradiography and immunoblotting analyses with the indicated antibodies were performed. **e**, In vitro kinase assays were performed by mixing purified His–PCK1 on Ni-NTA agarose beads with purified active GST–AKT1 in the presence of ATP for 1 h. His–PCK1-conjugated Ni-NTA agarose beads were washed with PBS five times and then incubated with SFB–INSIG1 or SFB–INSIG2 purified from Huh7 cells in the presence of [γ -³²P]ATP or [γ -³²P]GTP for 1 h. Autoradiography and immunoblotting analyses with the indicated antibodies were performed. **f**, An in vitro kinase assay was performed as in **e**, except that the His–PCK1-conjugated beads were incubated with SFB–INSIG1 purified from Huh7 cells in the presence of GTP for 1 h. Mass spectrometric analysis of a tryptic fragment at *m/z* 764.40387 Da (–1.96 mmu/–2.57 ppm), which was matched with the +2 charged peptide 205-RSRSLGLGITIAF-218, suggested that INSIG1 Ser207 was phosphorylated. The Mascot score was 53, and the expectation

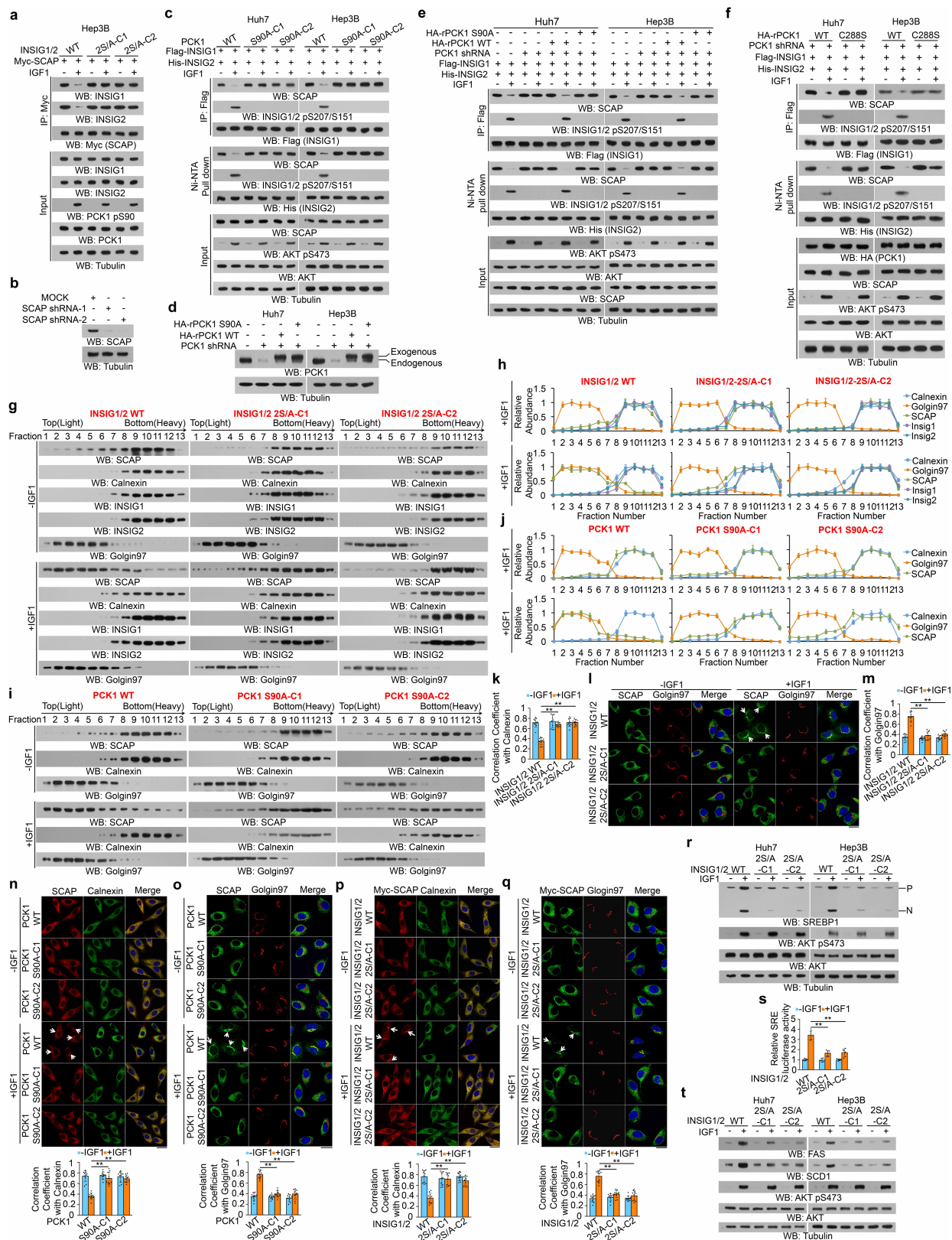
value was 0.021. **g**, PCK1-mediated phosphorylation residues in the cytosolic loop 2 of both INSIG1 and INSIG2. **h**, Alignment of protein sequences spanning INSIG1 Ser207 and INSIG2 Ser151 from different species. **i**, Bacterially purified wild-type His–PCK1 on Ni-NTA agarose beads was incubated with or without purified active GST–AKT1 in the presence of ATP for 1 h for an in vitro AKT kinase assay. The beads were then washed with PBS five times and incubated with or without wild-type SFB–INSIG2 or SFB–INSIG2(S151A) in the presence of [γ -³²P]GTP. Autoradiography and immunoblotting analyses with the indicated antibodies were performed. **j**, IHC analyses of human HCC samples were performed with the indicated antibodies in the presence or absence of a blocking peptide for INSIG1(pS207) and INSIG2(pS151). **k**, Huh7 cells expressing Flag–INSIG1 (top) or Flag–INSIG2 (bottom) were treated with or without IGF1 (100 ng ml⁻¹) for 1 h. Immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies in the presence or absence of a blocking peptide for INSIG1(pS207) and INSIG2(pS151). **l**, Enzyme kinetics plots of velocity relative to GTP concentration between purified wild-type His–PCK1 and His–PCK1(S90E). The *V*_{max} and *K*_m of PCK1 in phosphorylating an INSIG1 peptide at Ser207 were calculated (*n* = 6). Data are mean \pm s.d. **m**, **n**, Bacterially purified wild-type His–PCK1, His–PCK1(S90E) or His–PCK1(S90D) on Ni-NTA agarose beads were incubated with wild-type SFB–INSIG1 or SFB–INSIG1(S207A) (**m**) or wild-type SFB–INSIG2 or SFB–INSIG2(S151A) (**n**) in the presence of [γ -³²P]GTP. Autoradiography and immunoblotting assays with the indicated antibodies were performed. **o**, Wild-type Flag–INSIG1 or Flag–INSIG1(S207A) (left) or wild-type Flag–INSIG2 or Flag–INSIG2(S151A) (right) were transfected into Huh7 cells. Huh7 cells were stimulated with or without IGF1 (100 ng ml⁻¹) for 1 h. Immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies. **p**, Wild-type Flag–INSIG1 or Flag–INSIG1(S207A) (left) or wild-type Flag–INSIG2 or Flag–INSIG2(S151A) (right) were co-transfected with or without HA–myr-AKT1 into Huh7 cells. Immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies. **q**, Parental Huh7 cells and the indicated clones of cells with knock-in expression of PCK1(S90A) were transfected with SFB–INSIG1 (left) or SFB–INSIG2 (right). These cells were then stimulated with or without IGF1 (100 ng ml⁻¹) for 1 h. All experiments were repeated three times independently with similar results.



Extended Data Fig. 4 | Generation of Huh7 and Hep3B cells with knock-in expression of both INSIG1(S207A) and INSIG2(S151A), and PCK1-mediated phosphorylation of INSIG1/2 reduces the binding of oxysterols to INSIG1/2.

a, b, Genomic DNA was extracted from two individual clones of parental Huh7 and Hep3B cells with knock-in expression of INSIG1(S207A) (**a**) and INSIG2(S151A) (**b**). PCR products amplified from the indicated DNA fragments were separated on an agarose gel. **c, d**, Sequencing of parental and two individual clones of parental Huh7 and Hep3B cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) double mutants. The red line indicates the sgRNA-targeting sequence. The black line indicates the PAM. Blue arrows indicate mutated nucleotides. A mutated amino acid and its wild-type counterpart are indicated by the solid red box. **e**, Top, Flag/His-tagged INSIG2 immunoprecipitated and purified from Huh7 cells was incubated with the indicated GST–PCK1 proteins with or without active GST–AKT1 in the presence of ATP and GTP for 1 h. Immunoblotting analyses were performed with the indicated antibodies. Bottom, INSIG2-conjugated Ni-NTA agarose beads were washed and incubated with 400 nM [³H]25-hydroxycholesterol. Specifically bound [³H]25-hydroxycholesterol was measured ($n = 6$). $**P < 0.001$ (two-tailed t -test); NS, not significant ($P = 0.875, 0.846, 0.969, 0.924$ (left to right)). **f, g**, Top, Flag/His-tagged wild-type INSIG1 or INSIG1(S207A) (**f**) or Flag/His-tagged wild-type INSIG2 or INSIG2(S151A) (**g**) purified from Huh7 cells were incubated with purified wild-type GST–PCK1, GST–PCK1(S90D) or GST–PCK1(S90E) in the presence of GTP for 1 h. Immunoblotting analyses were performed with the indicated antibodies. Bottom, the INSIG1 or INSIG2 proteins on Ni-NTA agarose beads were washed with PBS five times and incubated with 400 nM [³H]25-hydroxycholesterol. Specifically bound [³H]25-hydroxycholesterol was measured ($n = 6$). $**P < 0.001$ (two-tailed t -test); NS, not significant ($P = 0.823, 0.445, 0.185$ (left to right) (**f**); $P = 0.320, 0.196, 0.735$ (left to right) (**g**)). **h**, Top,

Flag/His-tagged wild-type INSIG2 or INSIG2(S151A) purified from Huh7 cells were incubated with purified wild-type GST–PCK1 with or without purified active GST–AKT1 in the presence of ATP and GTP for 1 h. Immunoblotting analyses were performed with the indicated antibodies. Bottom, the INSIG2 protein on Ni-NTA agarose beads was washed with PBS five times and incubated with 400 nM [³H]25-hydroxycholesterol. Specifically bound [³H]25-hydroxycholesterol was measured ($n = 6$). $**P < 0.001$ (two-tailed t -test); NS, not significant ($P = 0.682, 0.947$ (left to right)). **i, j**, Flag/His-tagged wild-type INSIG1 or INSIG1(S207A) (**i**) or wild-type INSIG2 or INSIG2(S151A) (**j**) were purified from Huh7 treated with or without IGF1 (100 ng ml⁻¹) for 12 h. Immunoblotting analyses were performed with the indicated antibodies (left). The INSIG1 or INSIG2 proteins on Ni-NTA agarose beads were incubated with the indicated concentration of [³H]25-hydroxycholesterol. Specifically bound [³H]25-hydroxycholesterol was measured ($n = 6$) (right). **k, l**, Top, Flag/His-tagged INSIG1 (**k**) or INSIG2 (**l**) was expressed in parental Huh7 cells and the Huh7 cells with knock-in expression of PCK1(S90A). After these cells were treated with or without IGF1 (100 ng ml⁻¹) for 12 h, immunoblotting analyses were performed with the indicated antibodies. Bottom, the immunoprecipitated and purified INSIG1 or INSIG2 was incubated with the indicated concentration of [³H]25-hydroxycholesterol. Specifically bound [³H]25-hydroxycholesterol was measured ($n = 6$). **m, n**, Top, the immunoprecipitated and purified Flag/His-tagged wild-type INSIG1 or INSIG1(S207E) (**m**) or Flag/His-tagged wild-type INSIG2 or INSIG2(S151E) (**n**) from Huh7 cells was incubated with the indicated concentration of [³H]25-hydroxycholesterol. Specifically bound [³H]25-hydroxycholesterol was measured ($n = 6$). $**P < 0.001$ (two-tailed t -test). Data in **e–n** are mean \pm s.d. All experiments were repeated three times independently with similar results.

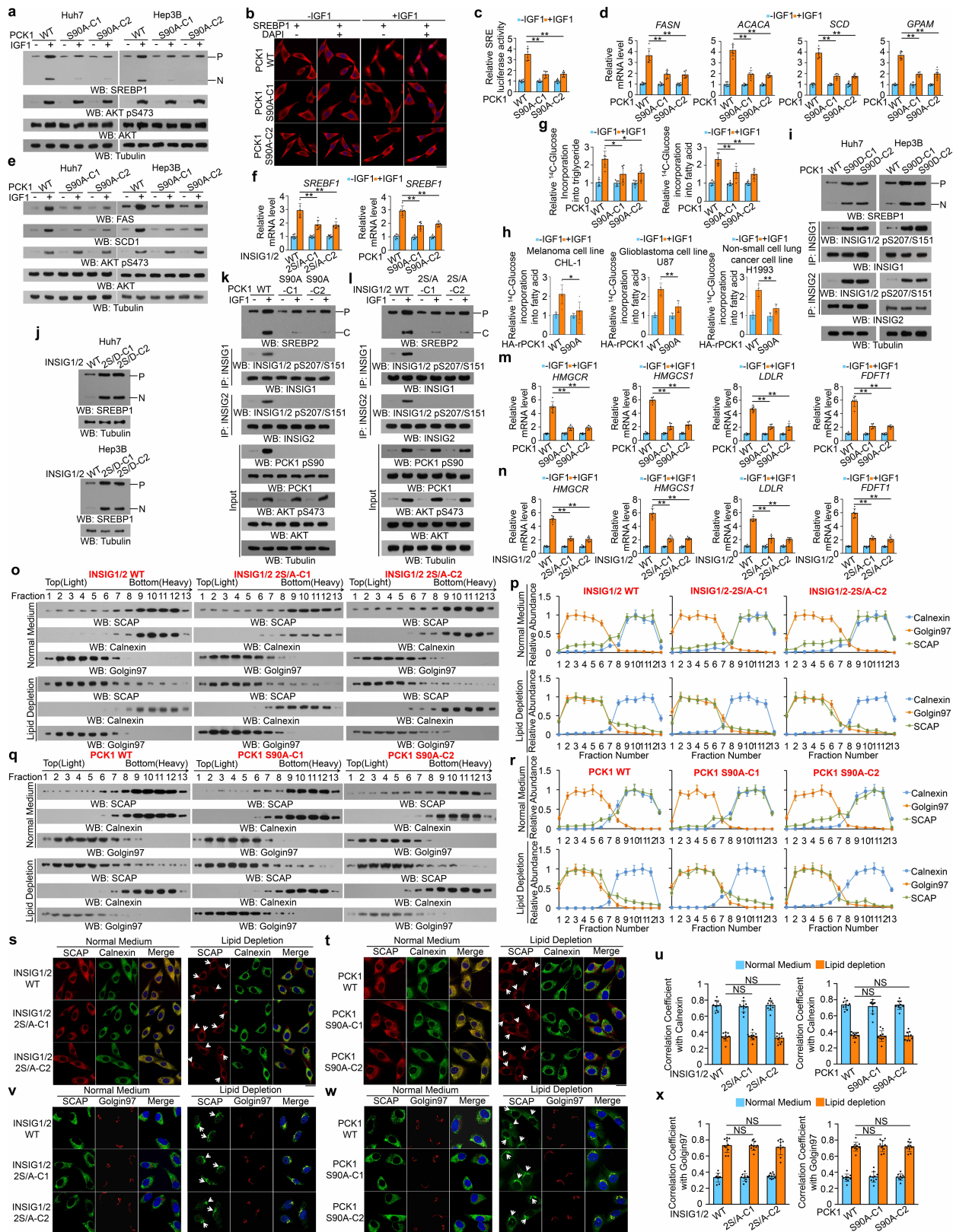


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | PCK1-mediated phosphorylation of INSIG1/2 promotes the translocation of SCAP from the ER to the Golgi apparatus.

a, Parental Hep3B cells and the indicated clones of Hep3B cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) double mutants were transfected with Myc-SCAP and stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. Immunoprecipitation and immunoblotting analyses with the indicated antibodies were performed. **b**, Validation of the specificity of the SCAP antibody. SCAP shRNA was expressed in 293T cells. Immunoblotting analyses were performed with the indicated antibodies. **c**, Parental Huh7 or Hep3B cells and the indicated clones with knock-in expression of PCK1(S90A) were transfected with the indicated plasmids and stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. Ni-NTA pull-down assays, immunoprecipitation and immunoblotting analyses with the indicated antibodies were performed. **d**, PCK1-depleted Huh7 (left) and Hep3B (right) cells were stably transfected with shRNA-resistant wild-type rPCK1 or rPCK1(S90A). Immunoblotting analyses were performed with the indicated antibodies. **e**, **f**, PCK1-depleted Huh7 (left) and Hep3B (right) cells with reconstituted expression of shRNA-resistant wild-type rPCK1 and rPCK1(S90A) (**e**) or rPCK1(C288S) (**f**) were transfected with vectors expressing Flag-INSIG1 and His-INSIG2. The cells were then stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. Ni-NTA pull-down assays, immunoprecipitation and immunoblotting analyses with the indicated antibodies were performed. **g**, **h**, Parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. The cells were then subjected to homogenization and cell fractionation using gradient centrifugation. Immunoblotting analyses were performed with the indicated antibodies (**g**). The relative distribution of each protein in different fractions was quantified by densitometric analysis of the blots ($n = 3$) (**h**). **i**, **j**, Parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of PCK1(S90A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. The cells were then subjected to homogenization and cell fractionation using gradient centrifugation. Immunoblotting analyses were performed with the indicated antibodies (**i**). The relative distribution of each protein in different fractions was quantified by densitometric analysis of the blots ($n = 3$) (**j**). **k**, Colocalization coefficients between SCAP and calnexin in the indicated cells. At least $n = 50$ cells from each independent experiment were analysed and representative data are shown. $^{**}P < 0.001$ (two-tailed t -test). **l**, Parental Huh7

cells and the indicated clones of Huh7 cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) were stimulated with or without IGF1 for 16 h. Immunofluorescence analyses were performed with the indicated antibodies. The white arrows indicate the Golgi-localized SCAP. Scale bar, 20 μ m. **m**, Colocalization coefficients between SCAP and golgin-97. At least $n = 50$ cells from each independent experiment were analysed and representative data are shown. $^{**}P < 0.001$ (two-tailed t -test). **n**, **o**, Top, parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of PCK1(S90A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. Immunofluorescence analyses were performed with the indicated antibodies. The white arrows indicate the Golgi-localized SCAP. Bottom, colocalization coefficients between SCAP and calnexin (**n**) or golgin-97 (**o**). At least $n = 50$ cells from each independent experiment were analysed and representative data are shown. $^{**}P < 0.001$ (two-tailed t -test). Scale bars, 20 μ m. **p**, **q**, Top, parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) double mutants were stably transfected with Myc-SCAP and stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. Immunofluorescence analyses were performed with the indicated antibodies. The white arrows indicate the Golgi-localized SCAP. Bottom, colocalization coefficients between Myc-SCAP and calnexin (**p**) or golgin-97 (**q**). At least $n = 50$ cells from each independent experiment were analysed and representative data are shown. $^{**}P < 0.001$ (two-tailed t -test). Scale bars, 20 μ m. **r**, Parental Huh7 and Hep3B cells and the indicated clones of these cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. N, N terminus of SREBP1; P, precursor of SREBP1. **s**, Parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) were transiently transfected with vectors expressing β -galactosidase and an SRE-driven luciferase reporter. Twenty-four hours later, the cells were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. The relative SRE luciferase activity after normalization to β -galactosidase activity is shown ($n = 6$). $^{**}P < 0.001$ (two-tailed t -test). **t**, Parental Huh7 and Hep3B cells and the indicated clones of these cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. Immunoblotting analyses were performed with the indicated antibodies. Data in **h**, **j**, **k**, **m**–**q**, **s** are mean \pm s.d. All experiments were repeated three times independently with similar results.



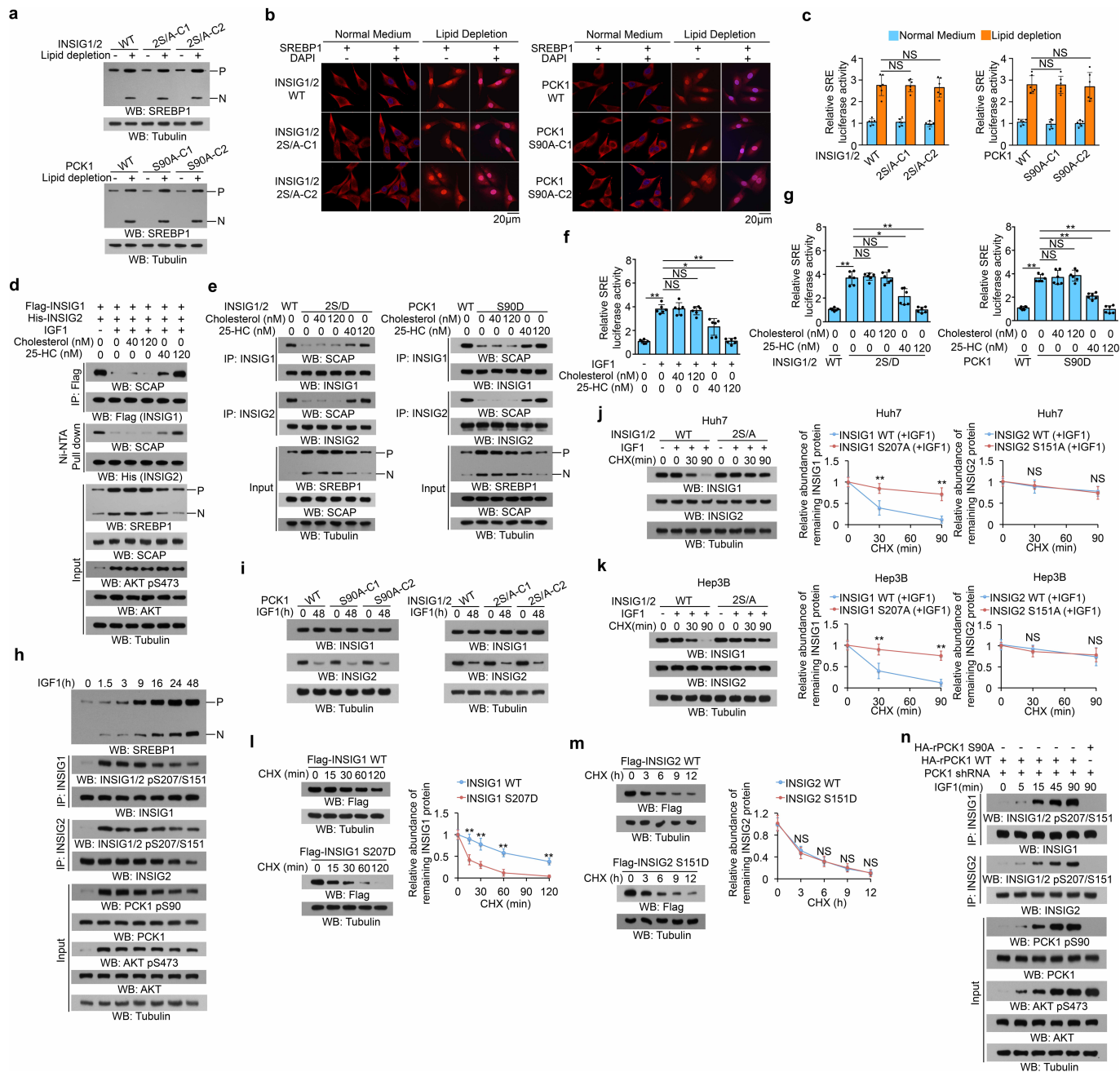
Extended Data Fig. 6 | See next page for caption.

Article

Extended Data Fig. 6 | PCK1-mediated INSIG1/2 phosphorylation is required for IGF1- induced SREBP activation for lipogenesis and does not affect the lipid depletion-induced translocation of SCAP from the ER to the Golgi apparatus.

a, Parental Huh7 (left) and Hep3B (right) cells and the indicated clones with knock-in expression of PCK1(S90A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. Immunoblotting analyses were performed as indicated. **b**, Parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of PCK1(S90A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. Immunofluorescence analyses were performed as indicated. Scale bar, 20 μm. **c**, Parental Huh7 cells and the indicated clones with knock-in expression of PCK1(S90A) were transiently transfected with vectors expressing β-galactosidase and an SRE-driven luciferase reporter and stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. The relative SRE luciferase activity after normalization to β-galactosidase activity is shown ($n = 6$). Data are mean ± s.d. ****** $P < 0.001$ (two-tailed t -test). **d**, Parental Huh7 cells and the indicated clones with knock-in expression of PCK1(S90A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. The mRNA expression levels for SREBP target genes were measured using quantitative PCR ($n = 6$). Data are mean ± s.d. ****** $P < 0.001$ (two-tailed t -test). **e**, Parental Huh7 (left) and Hep3B (right) cells and the indicated clones with knock-in expression of PCK1(S90A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. Immunoblotting analyses were performed as indicated. **f**, Parental Huh7 cells and the indicated clones with knock-in expression of INSIG1(S207A)/INSIG2(S151A) double mutants (left) or PCK1(S90A) (right) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. The mRNA expression levels for *SREBF1* were measured using quantitative PCR ($n = 6$). Data are mean ± s.d. *** $P = 0.002$, ****** $P < 0.001$** (two-tailed t -test). **g**, Parental Huh7 cells and the indicated clones with knock-in expression of PCK1(S90A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. The incorporation of ¹⁴C-glucose into triglycerides (left) and fatty acids (right) was measured ($n = 6$). Data are mean ± s.d. and were compared between groups using a two-tailed t -test. *** $P = 0.014$, 0.012** (left to right); **** $P = 0.004$, 0.001** (left to right). **h**, Endogenous PCK1-depleted CHL-1 human melanoma cells, U87 human glioblastoma cells and H1993 human non-small-cell lung cancer cells with reconstituted expression of shRNA-resistant wild-type HA-rPCK1 or HA-rPCK1(S90A) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. The incorporation of ¹⁴C-glucose into fatty acids was measured ($n = 6$). Data are mean ± s.d. and were

compared between groups using a two-tailed t -test. *** $P = 0.011$, ****** $P < 0.001$** . **i, j**, Parental Huh7 or Hep3B cells and the indicated clones with expression of PCK1(S90D) (**i**) or INSIG1(S207D)/INSIG2(S151D) double mutants (2S/D) (**j**) were collected for immunoblotting analyses as indicated. **k, l**, Parental Huh7 cells and the indicated clones with knock-in expression of PCK1(S90A) (**k**) or INSIG1(S207A)/INSIG2(S151A) (**l**) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. Immunoblotting analyses were performed as indicated. **m, n**, Parental Huh7 cells and the indicated clones with knock-in expression of PCK1(S90A) (**m**) or INSIG1(S207A)/INSIG2(S151A) (**n**) were stimulated with or without IGF1 (100 ng ml⁻¹) for 16 h. The mRNA expression levels for SREBP2 target genes were measured using quantitative PCR ($n = 6$). Data are mean ± s.d. **** $P < 0.001$** (two-tailed t -test). **o, p**, Parental Huh7 cells and the indicated clones with knock-in expression of INSIG1(S207A)/INSIG2(S151A) double mutants were incubated with a lipid-depleted medium and treated with lovastatin (an inhibitor of HMGCR) to inhibit cholesterol synthesis for 16 h. The cells were subjected to homogenization and cell fractionation using gradient centrifugation. Immunoblotting analyses were performed as indicated (**o**). The relative distribution of each protein in different fractions was quantified by densitometric analysis of the blots ($n = 3$) (**p**). Data are mean ± s.d. **q, r**, Parental Huh7 cells and the indicated clones with knock-in expression of PCK1(S90A) were incubated with or without lipid-depleted medium for 16 h. The cells were subjected to homogenization and cell fractionation using gradient centrifugation. Immunoblotting analyses were performed as indicated (**q**). The relative distribution of each protein in different fractions was quantified by densitometric analysis of the blots ($n = 3$) (**r**). Data are mean ± s.d. **s–x**, Parental Huh7 cells and the indicated clones with knock-in expression of INSIG1(S207A)/INSIG2(S151A) double mutants (**s, v**) or PCK1(S90A) (**t, w**) were incubated with or without lipid-depleted medium for 16 h. Immunofluorescence analyses were performed as indicated. Scale bars, 20 μm. The white arrows indicate the Golgi-localized SCAP. The colocalization coefficients between SCAP and the ER marker calnexin (**u**) and the Golgi apparatus marker golgin-97 (**x**) are shown. At least $n = 50$ cells from each independent experiment were analysed and representative data are shown. Data are mean ± s.d. and were compared between groups using a two-tailed t -test. NS, not significant ($P = 0.788$, 0.514, 0.689, 0.693 (left to right) (**u**); $P = 0.976$, 0.606, 0.750, 0.940 (left to right) (**x**)). All experiments were repeated three times independently with similar results.

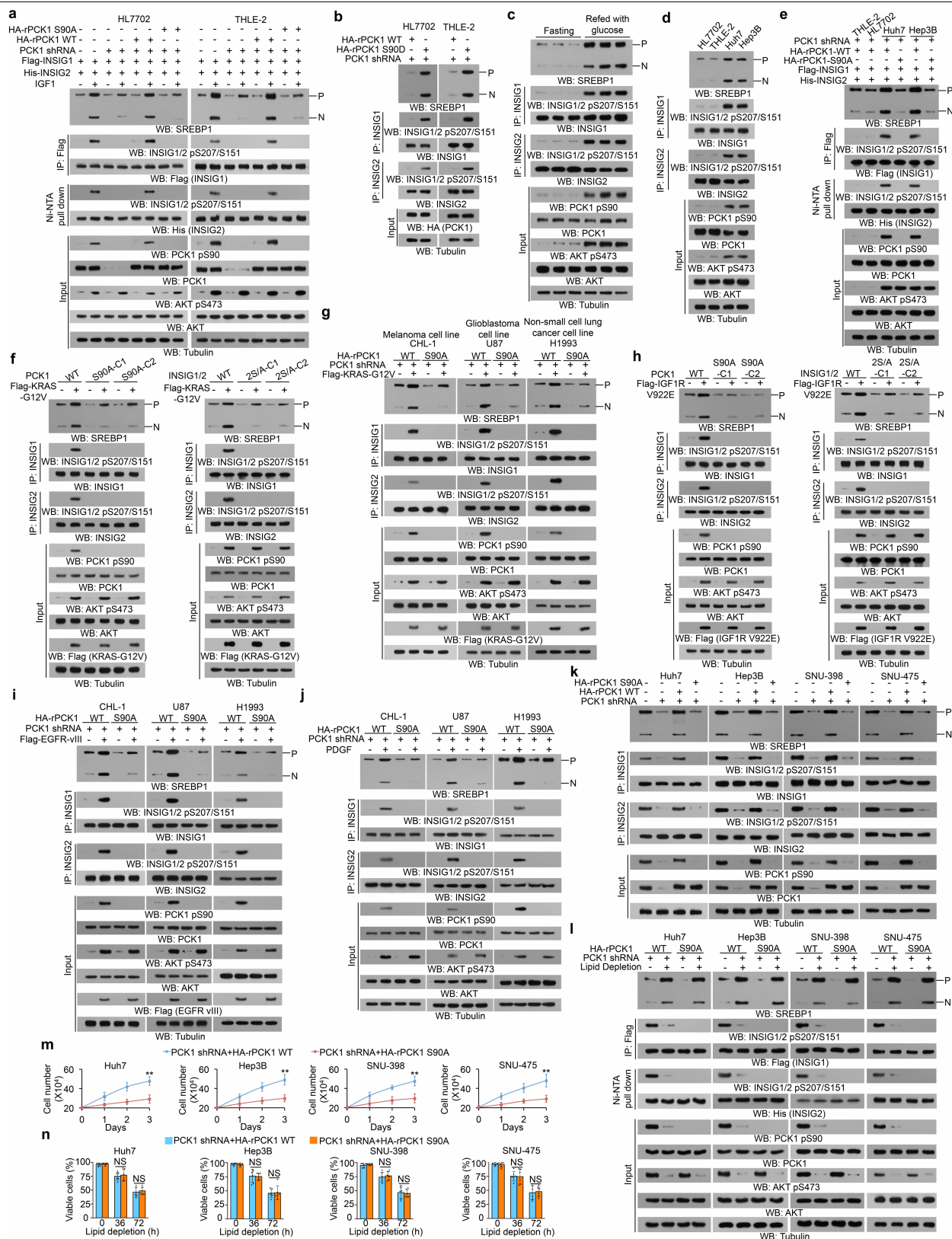


Extended Data Fig. 7 | See next page for caption.

Article

Extended Data Fig. 7 | PCK1-mediated phosphorylation of INSIG1/2, which does not affect lipid-depletion-induced activity of SREBP1, promotes rapid activation of SREBP1 and degradation of INSIG1. **a**, Parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) double mutants (top) or knock-in expression of PCK1(S90A) (bottom) were incubated with or without lipid-depleted medium for 16 h. Immunoblotting analyses were performed with the indicated antibodies. **b**, Parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) double mutants (left) or knock-in expression of PCK1(S90A) (right) were incubated with or without lipid-depleted medium for 16 h. Immunofluorescence analyses were performed with the indicated antibodies. **c**, Parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of INSIG1(S207A)/INSIG2(S151A) double mutants (left) or knock-in expression of PCK1(S90A) (right) were transiently transfected with vectors expressing β -galactosidase and an SRE-driven luciferase reporter. Luciferase activity was determined after the cells were incubated with or without lipid-depleted medium for 16 h. The relative SRE luciferase activity after normalization to β -galactosidase activity is shown ($n = 6$). NS, not significant ($P = 0.965, 0.699, 0.967, 0.767$ (left to right)). **d**, Huh7 cells expressing Flag-INSIG1 and His-INSIG2 were stimulated with or without IGF1 (100 ng ml^{-1}) for 16 h in the presence of 40 nM or 120 nM cholesterol or 25-hydroxycholesterol. Ni-NTA pull-down assays, immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies. 25-HC, 25-hydroxycholesterol. **e**, Parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of both INSIG1(S207D) and INSIG2(S151D) (left) or knock-in expression of PCK1(S90D) (right) were incubated with 40 nM or 120 nM cholesterol or 25-hydroxycholesterol for 16 h. Immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies. **f**, Huh7 cells were transiently transfected with vectors expressing β -galactosidase and an SRE-driven luciferase reporter. Twenty-four hours later, the cells were stimulated with or without IGF1 (100 ng ml^{-1}) for 16 h in the presence of 40 nM or 120 nM cholesterol or 25-hydroxycholesterol. The relative SRE luciferase activity after normalization to β -galactosidase activity is shown ($n = 6$). $^*P = 0.001$, $^{**}P < 0.001$ (two-tailed t -test); NS, not significant ($P = 0.921, 0.579$ (left to right)). **g**, Parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of INSIG1(S207D)/INSIG2(S151D) double mutants (left)

or knock-in expression of PCK1(S90D) (right) were transiently transfected with vectors expressing β -galactosidase and an SRE-driven luciferase reporter. Luciferase activity was determined after the cells were incubated with 40 nM or 120 nM cholesterol or 25-hydroxycholesterol for 16 h. The relative SRE luciferase activity after normalization to β -galactosidase activity is shown ($n = 6$). $^*P = 0.002$, $^{**}P < 0.001$ (two-tailed t -test); NS, not significant ($P = 0.642, 0.957, 0.842, 0.372$ (left to right)). **h**, Huh7 cells were treated with IGF1 (100 ng ml^{-1}) for the indicated time periods. Immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies. **i**, Parental Huh7 cells and the indicated clones of Huh7 cells with knock-in expression of PCK1(S90A) (left) or INSIG1(S207A)/INSIG2(S151A) double mutants (right) were treated with or without IGF1 (100 ng ml^{-1}) for 48 h. Immunoblotting analyses were performed with the indicated antibodies. **j, k**, Serum-starved parental Huh7 (**j**) and Hep3B (**k**) cells or the indicated clones with knock-in expression of INSIG1(S207A)/INSIG2(S151A) double mutants were stimulated with or without IGF1 (100 ng ml^{-1}) in the presence or absence of CHX ($100 \mu\text{g ml}^{-1}$) for the indicated time periods. Immunoblotting analyses were performed with the indicated antibodies (left). The relative abundance of remaining INSIG1 or INSIG2 protein was quantified (right) ($n = 6$). $^{**}P < 0.001$ (two-tailed t -test); NS, not significant ($P = 0.721, 0.637$ (left to right) (**j**); $P = 0.310, 0.853$ (left to right) (**k**)). **l**, Huh7 cells expressing Flag-tagged wild-type INSIG1 (top) or INSIG1(S207D) (bottom) were treated with CHX ($100 \mu\text{g ml}^{-1}$) for the indicated time periods. Immunoblotting analyses were performed with the indicated antibodies (left). The relative abundance of the remaining INSIG1 protein was quantified (right) ($n = 6$). $^{**}P < 0.001$ (two-tailed t -test). **m**, Huh7 cells expressing Flag-tagged wild-type INSIG2 (top) or INSIG2(S151D) (bottom) were treated with CHX ($100 \mu\text{g ml}^{-1}$) for the indicated time periods. Immunoblotting analyses were performed with the indicated antibodies (left). The relative abundance of the remaining INSIG2 protein was quantified (right) ($n = 6$). $^{**}P < 0.001$ (two-tailed t -test); NS, not significant ($P = 0.395, 0.973, 0.636, 0.882$ (left to right)). **n**, Endogenous PCK1-depleted Huh7 cells with reconstituted expression of shRNA-resistant wild-type PCK1 or rPCK1(S90A) were treated with IGF1 (100 ng ml^{-1}) for the indicated time periods. Immunoprecipitation and immunoblotting analyses were performed with the indicated antibodies. Data in **c, f, g, j–m** are mean \pm s.d. All experiments were repeated three times independently with similar results.

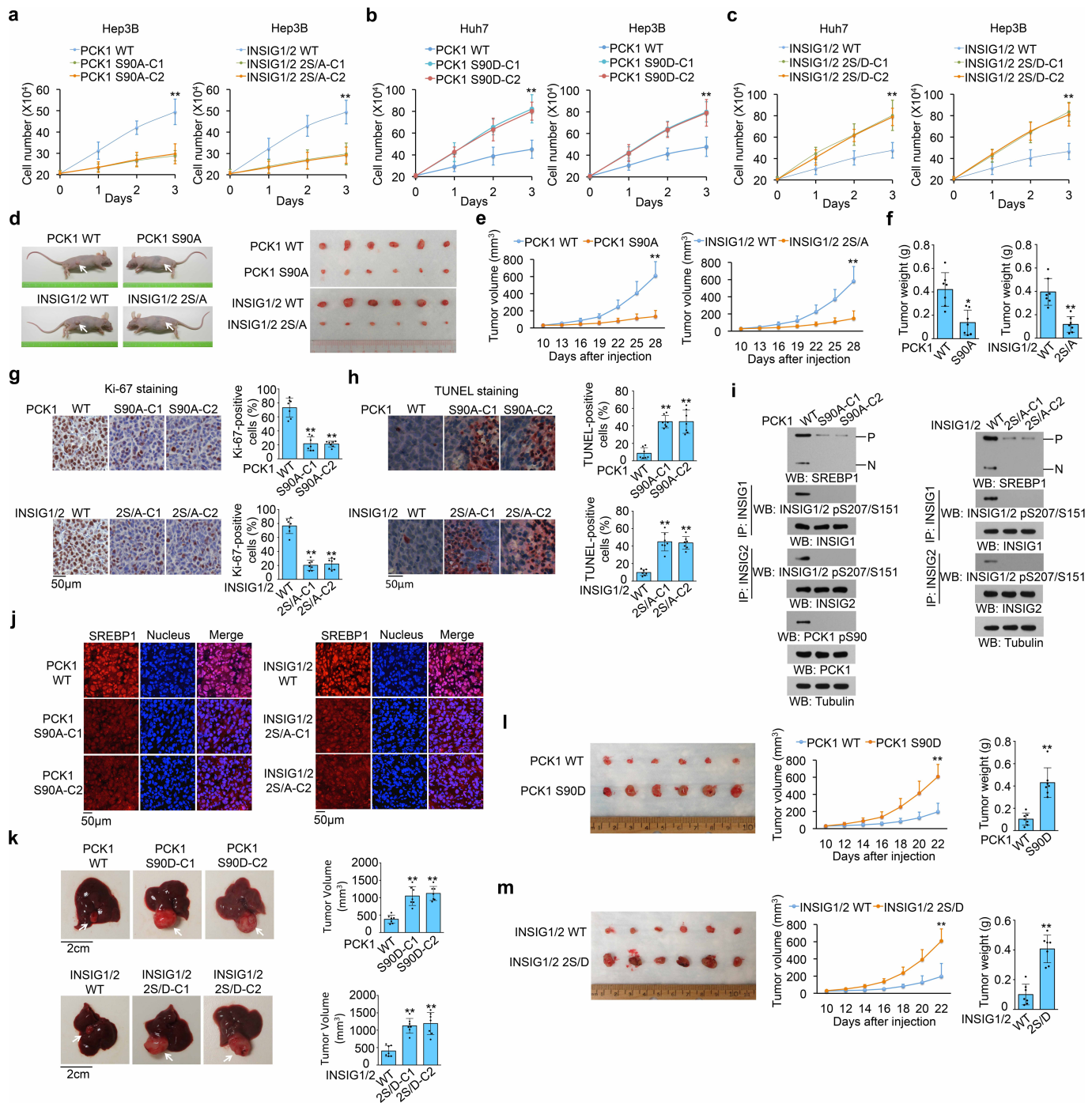


Extended Data Fig. 8 | See next page for caption.

Article

Extended Data Fig. 8 | PCK1-mediated phosphorylation of INSIG1 Ser207 and INSIG2 Ser151 and SREBP1 activation are induced by oncogene- or growth-factor-mediated activation of AKT in several cancer types. a, b, e–l. Cells were transfected with the indicated plasmids and immunoprecipitation or immunoblotting analyses were performed with the indicated antibodies. **a**, Endogenous PCK1-depleted HL7702 (left) and THLE-2 (right) cells with reconstituted expression of shRNA resistant wild-type HA-rPCK1 or HA-rPCK1(S90A) were treated with or without IGF1 (100 ng ml⁻¹) for 16 h. **b**, Endogenous PCK1-depleted HL7702 (left) and THLE-2 (right) cells with reconstituted expression of shRNA resistant wild-type HA-rPCK1 or HA-rPCK1(S90D) were analysed. **c**, C57BL/6J male mice (eight-week-old) fasted for 24 h were intraperitoneally injected with or without glucose (1 g per kg body weight). After 6 h, the mouse livers were dissected for immunoprecipitation and immunoblotting analyses. **d**, HL7702, THLE-2, Huh7 and Hep3B cells were analysed by Ni-NTA pull-down, immunoprecipitation and immunoblotting assays. **e**, Endogenous PCK1-depleted HL7702, THLE-2, Huh7 and Hep3B cells with reconstituted expression of shRNA resistant wild-type HA-rPCK1 or HA-rPCK1(S90A) were analysed. **f**, Parental Huh7 cells and the indicated clones with knock-in expression of PCK1(S90A) (left) or INSIG1(S207A)/INSIG2(S151A) double mutants (right) were transfected with or without Flag-KRAS(G12V). **g**, Endogenous PCK1-depleted CHL-1, U87 and H1993 cells with reconstituted expression of shRNA resistant wild-type HA-rPCK1 or HA-rPCK1(S90A) were transfected with the indicated plasmids. **h**, Parental Huh7 cells and the

indicated clones with knock-in expression of PCK1(S90A) (left) or INSIG1(S207A)/INSIG2(S151A) double mutants (right) were transfected with or without constitutively active Flag-IGF1R(V922E). **i**, Endogenous PCK1-depleted CHL-1, U87 and H1993 cells with reconstituted expression of shRNA resistant wild-type HA-rPCK1 or HA-rPCK1(S90A) were transfected with the indicated plasmids. **j**, Endogenous PCK1-depleted CHL-1, U87 and H1993 cells with reconstituted expression of shRNA resistant wild-type HA-rPCK1 or HA-rPCK1(S90A) were transfected with the indicated plasmids and treated with or without PDGF (30 ng ml⁻¹) for 16 h. **k, m**, Endogenous PCK1-depleted Huh7, Hep3B, SNU-398 and SNU-475 cells with reconstituted expression of shRNA resistant wild-type HA-rPCK1 or HA-rPCK1(S90A) were analysed by immunoprecipitation and immunoblotting analyses with the indicated antibodies (**k**). The cells were plated and then collected and counted for 3 days ($n = 6$) (**m**). Data are mean \pm s.d. ****** $P < 0.001$ (two-tailed t -test). **l, n**, Endogenous PCK1-depleted Huh7, Hep3B, SNU-398 and SNU-475 cells with reconstituted expression of shRNA resistant wild-type HA-rPCK1 or HA-rPCK1(S90A) were transfected with Flag-INSIG1 and His-INSIG2. After incubation with or without lipid-depleted medium for 16 h, the cells were collected for Ni-NTA pull-down, immunoprecipitation and immunoblotting analyses as indicated (**l**). Viable cells were counted for 3 days after lipid depletion ($n = 6$) (**n**). Data are mean \pm s.d. NS, not significant ($P = 0.708, 0.619, 0.888, 0.901, 0.633, 0.788, 0.902, 0.764$ (left to right)). All experiments were repeated three times independently with similar results.

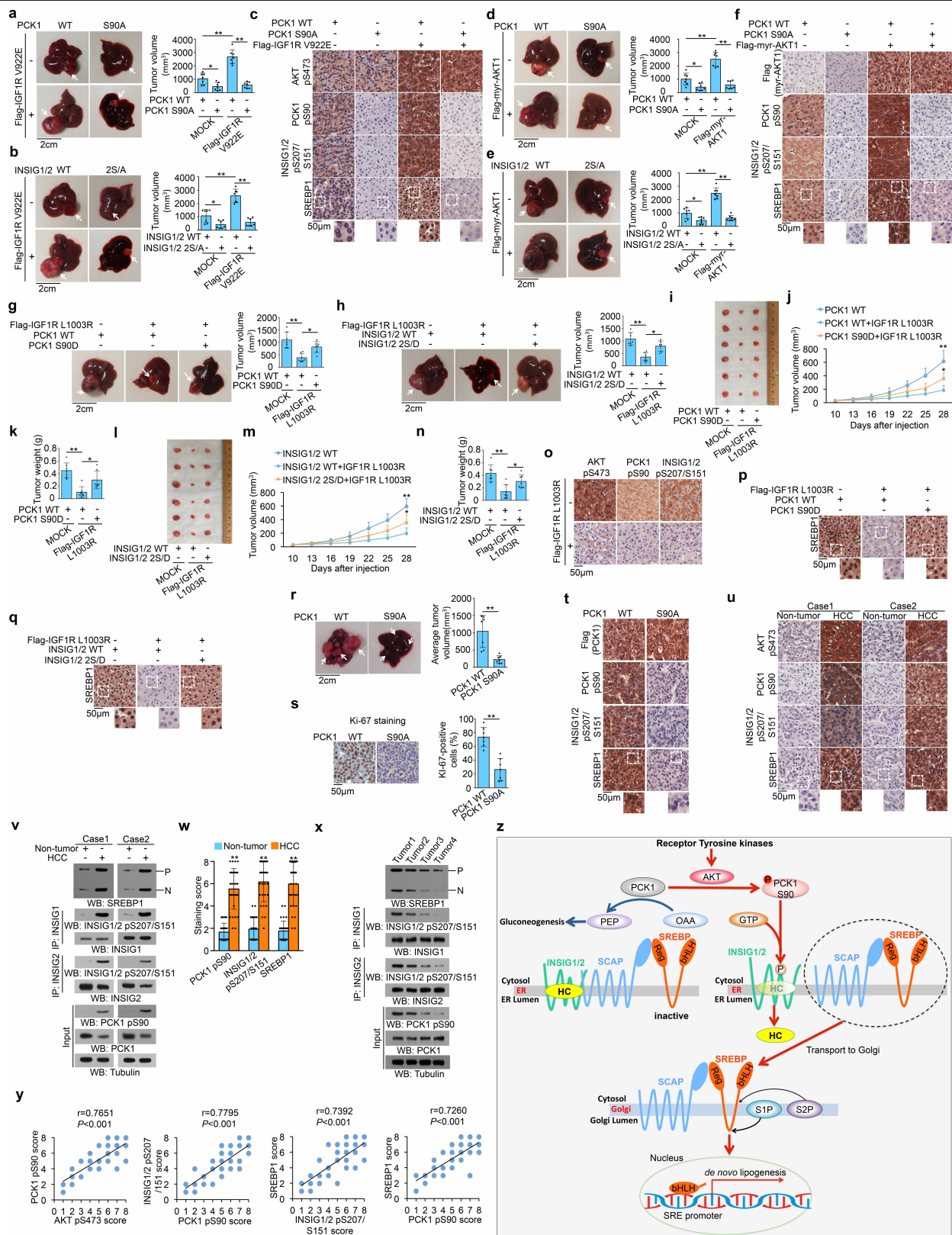


Extended Data Fig. 9 | See next page for caption.

Article

Extended Data Fig. 9 | PCK1-mediated phosphorylation of INSIG1/2 promotes the proliferation of liver cancer cells and the growth of tumours in mice. **a**, Parental Hep3B cells (2×10^5) and the indicated clones of Hep3B cells (2×10^5) with knock-in expression of PCK1(S90A) (left) or INSIG1(S207A)/INSIG2(S151A) double mutants (right) were plated for 3 days. The cells were then collected and counted. Data are mean \pm s.d. ($n = 6$). $**P < 0.001$ (two-tailed t -test). **b, c**, Parental Huh7 and Hep3B cells (2×10^5) and the indicated clones of Huh7 and Hep3B cells with knock-in expression of PCK1(S90D) (**b**) or INSIG1(S207D)/INSIG2(S151D) double mutants (**c**) were plated for 3 days. The cells were then collected and counted. Data are mean \pm s.d. ($n = 6$). $**P < 0.001$ (two-tailed t -test). **d–h**, Parental Huh7 cells (1×10^6) or the clones with knock-in expression of PCK1(S90A) or INSIG1(S207A)/INSIG2(S151A) double mutants were subcutaneously injected into the left or right flanks of athymic nude mice, respectively ($n = 7$ per group) (**d**, left). The resulting tumours were resected 28 days after injection (**d**, right). The growth of xenografted tumours in the mice was measured (**e**) and the tumours were weighed (**f**). Data are mean \pm s.d. ($n = 7$). $*P = 0.002$, $**P < 0.001$ compared with the wild-type group (two-tailed t -test). **g**, IHC analyses of tumour samples were performed with an anti-Ki67 antibody. Ki67-positive cells were quantified (right). **h**, TUNEL analyses of the indicated tumour samples were performed. Apoptotic cells were stained brown and quantified in $n = 10$ microscopic fields (right). Data are mean \pm s.d. $**P < 0.001$

compared with the wild-type group (two-tailed t -test). **i, j**, Huh7 cells (1×10^6) with or without knock-in expression of PCK1(S90A) or INSIG1(S207A)/INSIG2(S151A) double mutants were intrahepatically injected into athymic nude mice ($n = 7$ per group). At 28 days after injection, the mice were euthanized and the liver tumours were dissected for immunoprecipitation, immunoblotting (**i**) and immunofluorescence (**j**) analyses with the indicated antibodies. **k**, Huh7 cells (1×10^6) with or without knock-in expression of PCK1(S90D) (top) or INSIG1(S207D)/INSIG2(S151D) double mutants (bottom) were intrahepatically injected into athymic nude mice ($n = 7$ per group). The mice were euthanized and examined for tumour growth 22 days after injection. The arrows indicate tumours. Tumour volumes were calculated (right). Data are mean \pm s.d. ($n = 7$). $**P < 0.001$ compared with the wild-type group (two-tailed t -test). **l, m**, Huh7 cells (1×10^6) and Huh7 cells with knock-in expression of PCK1(S90D) (**l**) or INSIG1(S207D)/INSIG2(S151D) double mutants (**m**) were subcutaneously injected into the left or right flanks of athymic nude mice, respectively ($n = 7$ per group). The resulting tumours were resected 22 days after injection (left). The growth of xenografted tumours in the mice was measured (middle) and the tumours were weighed (right). Data are mean \pm s.d. ($n = 7$). $**P < 0.001$ compared with the wild-type group (two-tailed t -test). All experiments were repeated three times independently with similar results.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Activation of the IGF1R-AKT-PCK1-INSIG1/2 signalling cascade is required for the growth of liver tumours and correlates with poor prognosis for HCC. **a, b**, Huh7 cells (1×10^6) with or without knock-in expression of PCK1(S90A) (**a**) or INSIG1(S207A)/INSIG2(S151A) double mutants (**b**) were stably transfected with or without Flag-IGF1R(V922E) and intrahepatically injected into athymic nude mice ($n = 7$ per group). The mice were euthanized and examined for tumour growth 28 days after injection. The arrows indicate tumours. Tumour volumes were calculated (right). Data are mean \pm s.d. ($n = 7$). $^*P = 0.011$ (**a**), 0.013 (**b**), $^{**}P < 0.001$ (two-tailed t -test). **c**, IHC analyses of xenografted tumours from nude mice were performed with the indicated antibodies. The regions in white boxes are shown at higher magnification below. **d, e**, Huh7 cells (1×10^6) with or without knock-in expression of PCK1(S90A) (**d**) or INSIG1(S207A)/INSIG2(S151A) double mutants (**e**) were stably transfected with or without Flag-myr-AKT1 and intrahepatically injected into athymic nude mice ($n = 7$ per group). The mice were euthanized and examined for tumour growth 28 days after injection. The arrows indicate tumours. Tumour volumes were calculated (right). Data are mean \pm s.d. ($n = 7$). $^*P = 0.015$ (**d**), 0.010 (**e**), $^{**}P < 0.001$ (two-tailed t -test). **f**, IHC analyses of xenografted tumours from nude mice were performed with the indicated antibodies. The regions in white boxes are shown at higher magnification below. **g, h**, Huh7 cells (1×10^6) with or without knock-in expression of PCK1(S90D) (**g**) or INSIG1(S207D)/INSIG2(S151D) double mutants (**h**) were stably transfected with Flag-IGF1R(L1003R) and intrahepatically injected into athymic nude mice ($n = 7$ per group). The mice were euthanized and examined for tumour growth 28 days after injection. The arrows indicate tumours. Tumour volumes were calculated (right). Data are mean \pm s.d. ($n = 7$). $^*P = 0.003$ (**g**), 0.005 (**h**), $^{**}P < 0.001$ (two-tailed t -test). **i–n**, Huh7 cells (1×10^6) with or without knock-in expression of PCK1(S90D) (**i–k**) or INSIG1(S207D)/INSIG2(S151D) double mutants (**l–n**) were stably transfected with or without Flag-IGF1R(L1003R) and subcutaneously injected into the left flanks of athymic nude mice ($n = 7$ per group). The resulting tumours were resected 28 days after injection (**i, l**). The growth of xenografted tumours in the mice was measured (**j, m**). The tumours were weighed (**k, n**). Data are mean \pm s.d. ($n = 7$). $^*P = 0.002$ (**j**), 0.011 (**k**), 0.009 (**m**), 0.016 (**n**), $^{**}P < 0.001$ compared with the PCK1 wild-type + IGF1R(L1003R) group (**j, k**) or compared with the INSIG1/2 wild-type + IGF1R(L1003R) group (**m, n**) (two-tailed t -test). **o–q**, Huh7 cells

(1×10^6) with or without knock-in expression of PCK1(S90D) or INSIG1(S207D)/INSIG2(S151D) double mutants were stably transfected with or without Flag-IGF1R(L1003R) and intrahepatically injected into athymic nude mice ($n = 7$ per group). The mice were euthanized 28 days after injection. IHC analyses of xenografted tumours from nude mice were performed with the indicated antibodies. The regions in white boxes are shown at higher magnification below. **r–t**, Wild-type pT3-EF1 α -Flag-PCK1 (or pT3-EF1 α -Flag-PCK1(S90A)), pT3-EF1 α -HA-myr-AKT1 and pT3-EF1 α -V5-c-Met, along with the sleeping beauty transposase (SB), were stably expressed in the mouse liver using hydrodynamic transfection into FVB/N mice ($n = 7$ per group). After 14 weeks, the mice were euthanized and representative liver tumours are shown (**r**, left). The average tumour volumes were measured (**r**, right) ($n = 7$ per group). $^{**}P < 0.001$ (two-tailed t -test). **s**, IHC analyses of tumour samples were performed with an anti-Ki67 antibody (left). Ki67-positive cells were quantified in 10 microscopic fields (right). $^{**}P < 0.001$ (two-tailed t -test). **t**, IHC analyses of liver tumours from nude mice were performed with the indicated antibodies. The regions in white boxes are shown at higher magnification below. **u**, IHC staining of 30 human HCC and matched non-tumour tissue samples was performed with the indicated antibodies. Representative images of two cases are shown. The regions in white boxes are shown at higher magnification below. **v**, Representative immunoblotting analyses of two cases of human HCC and matched non-tumour tissue samples was performed with the indicated antibodies. **w**, The indicated staining scores for PCK1(pS90), INSIG1(pS207)/INSIG2(pS151) and SREBP1 expression levels in HCC and matched non-tumour liver samples ($n = 30$) were compared using a paired t -test (two-tailed). Data are mean \pm s.d. $^{**}P < 0.001$ compared with the non-tumour adjacent tissue. **x**, Representative immunoblotting analyses of four different human HCC samples were performed with the indicated antibodies. **y**, IHC staining of human HCC samples with the indicated antibodies was scored, and correlation analyses were performed. A Pearson correlation test was used (two-tailed) ($n = 40$). Note that the scores for some samples overlap. **z**, Mechanism for PCK1-promoted activation of SREBP1 and lipogenesis. PEP, phosphoenolpyruvate; OAA, oxaloacetate; HC, hydroxycholesterol; bHLH, basic helix-loop-helix protein; Reg, regulatory domain of SREBP. All experiments were repeated three times independently with similar results.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☐ ☒ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☐ ☒ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated
- ☐ ☒ Clearly defined error bars
State explicitly what error bars represent (e.g. SD, SE, CI)

Our web collection on [statistics for biologists](#) may be useful.

Software and code

Policy information about [availability of computer code](#)

Data collection

Image J (version 1.8.0) has been used for the quantification of immunoblotting and Immunofluorescence staining analyses. FV10-ASW Viewer software (Version 4.2b) is used to collect the immunofluorescence data. Mascot software program and Proteome Discoverer software program (Version 2.2) has been used for the proteomics studies.

Data analysis

Excel (2007) has been used for the two-sided T-test. Realtime PCR data: ABI 7500 software (version 2.3). GraphPad Prism (V.8.1.1) has been used for patients' survival analyses.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/authors/policies/ReportingSummary-flat.pdf](https://www.nature.com/authors/policies/ReportingSummary-flat.pdf)

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The group sizes of the animals chosen are based on the numbers we used for previous publications, which is most optimal to generate statistically significant results. Triplicated experiments were conducted so that statistical significance can be tested.
Data exclusions	No data were excluded.
Replication	Experiments were repeated in triplication. All replication were successful.
Randomization	The samples/cells for IHC studies were randomized to be examined (No specific methods were used).
Blinding	Blinding is not relevant because all experiment groups were conducted with different treatments.

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).
Research sample	State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.
Sampling strategy	Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.
Data collection	Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.
Timing	Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Non-participation	State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.
Randomization	If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Briefly describe the study. For quantitative data include treatment factors and interactions, design structure (e.g. factorial, nested, hierarchical), nature and number of experimental units and replicates.
Research sample	Describe the research sample (e.g. a group of tagged <i>Passer domesticus</i> , all <i>Stenocereus thurberi</i> within Organ Pipe Cactus National Monument), and provide a rationale for the sample choice. When relevant, describe the organism taxa, source, sex, age range and any manipulations. State what population the sample is meant to represent when applicable. For studies involving existing datasets, describe the data and its source.
Sampling strategy	Note the sampling procedure. Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.
Data collection	Describe the data collection procedure, including who recorded the data and how.
Timing and spatial scale	Indicate the start and stop dates of data collection, noting the frequency and periodicity of sampling and providing a rationale for these choices. If there is a gap between collection periods, state the dates for each sample cohort. Specify the spatial scale from which the data are taken
Data exclusions	If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.
Reproducibility	Describe the measures taken to verify the reproducibility of experimental findings. For each experiment, note whether any attempts to repeat the experiment failed OR state that all attempts to repeat the experiment were successful.
Randomization	Describe how samples/organisms/participants were allocated into groups. If allocation was not random, describe how covariates were controlled. If this is not relevant to your study, explain why.
Blinding	Describe the extent of blinding used during data acquisition and analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.
Did the study involve field work?	<input type="checkbox"/> Yes <input type="checkbox"/> No

Field work, collection and transport

Field conditions	Describe the study conditions for field work, providing relevant parameters (e.g. temperature, rainfall).
Location	State the location of the sampling or experiment, providing relevant parameters (e.g. latitude and longitude, elevation, water depth).
Access and import/export	Describe the efforts you have made to access habitats and to collect and import/export your samples in a responsible manner and in compliance with local, national and international laws, noting any permits that were obtained (give the name of the issuing authority, the date of issue, and any identifying information).
Disturbance	Describe any disturbance caused by the study and how it was minimized.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Unique biological materials
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input type="checkbox"/>	<input checked="" type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Unique biological materials

Policy information about [availability of materials](#)

Obtaining unique materials

Describe any restrictions on the availability of unique materials OR confirm that all unique materials used are readily available from the authors or from standard commercial sources (and specify these sources).

Antibodies

Antibodies used

WB: GST (Santa Cruz Biotechnology, sc-138, B-14, 1:2000)
 WB: Tubulin (Santa Cruz Biotechnology, sc-8035, TU-02, 1:1000),
 WB: ERK1/2 (Santa Cruz Biotechnology, sc-514302, C-9, 1:2000)
 WB and IF: Insig1 (Proteintech, 22115-1-AP, Ag17420, WB: 1:1000; IF: 1:200),
 WB: Insig2 antibody (Thermo Fisher, PA5-41707, Rabbit polyclonal, WB: 1:1000),
 WB: Insig1 antibody (Thermo Fisher, PA5-97876, Rabbit polyclonal, WB: 1:1000),
 WB: AKT pS473 (Cell signaling, #4060, D9E, WB: 1:1000)
 WB: AKT (Cell signaling, #9272, Rabbit Polyclonal, WB: 1:1000)
 WB: SRC (Cell signaling, #2108, Rabbit Polyclonal, WB: 1:1000)
 WB: Phospho-p44/42 MAPK (Erk1/2) (Thr202/Tyr204) (Cell signaling, #9101, Rabbit, WB: 1:1000)
 WB: c-Jun (Cell signaling, #9165, 60A8, WB: 1:1000)
 WB: c-Jun pS73 (Cell signaling, #3270, D47G9, WB: 1:1000)
 WB: PCK1 (Cell signaling, #12940, D12F5, WB: 1:1000)
 WB: Flag (Sigma-Aldrich, F1804, M2, Mouse Monoclonal, WB: 1:5000)
 WB: Flag (Sigma-Aldrich, F7425, Rabbit Polyclonal, WB: 1:5000)
 WB: His (Sigma-Aldrich, SAB1305538, 6AT18, WB: 1:5000)
 WB: Src (phospho Y418) (ABCAM, ab4816, Rabbit polyclonal, WB: 1:1000)
 WB and IP: Insig1 (Santa Cruz Biotechnology, sc-390504, Mouse Monoclonal, A-9, WB: 1:2000; IP: 500 µg/ 0.25 ml agarose in 1 ml)
 WB and IP: PCK1 (Proteintech, 16754-1-AP, Ag10261, WB: 1:1000; IP: 0.5-4.0 µg)
 WB and IP: Insig2 (Proteintech, 24766-1-AP, Ag14072, WB: 1:1000; IP: 0.5-4.0 µg)
 WB and IP: Myc tag (ABCAM, ab9106, Rabbit polyclonal, WB: 1:2000, IP:1:500)
 WB and IP: HA (Cell signaling, #3724, C29F4, WB: 1:1000, IP: 1:250)
 WB, IP and IF: Myc-tag mouse (Cell signaling, #2276, 9B11, WB: 1:2000, IP:1:500, IF: 1:500)
 WB and IF: PCK1 (Novus, H00005105-M1, 3E4, Mouse Monoclonal, WB: 1:1000; IF: 1:200)
 WB and IF: Calnexin (ABCAM, ab22595, Rabbit polyclonal, WB: 1:1000; IF:1:200)
 WB and IF: Calnexin (Thermo Fisher, MA3-027, AF18, Mouse Monoclonal, WB: 1:1000; IF:1:200)
 WB and IF: SCAP antibody (ABCAM, ab190103, Rabbit polyclonal, WB: 1:1000; IF: 1:300)
 WB: Golgi 97 (ABCAM, ab84340, Rabbit polyclonal, WB: 1:1000)
 WB: SCD1 (ABCAM, ab19862, CD.E10, Mouse Monoclonal, WB: 1:1000)
 WB: FASN (ABCAM, ab22759, Rabbit polyclonal, WB: 1:1000)
 WB: SREBP1 (BD Biosciences, 557036, IgG 2A4, WB: 1:1000)
 WB: SREBP2 (BD Biosciences, 557037, IgG-1C6, WB: 1:1000)
 WB: anti-Rabbit IgG heavy chain (HRP) (ABCAM, ab99702, 2A9, Mouse, WB: 1:5000)
 WB: Rabbit IgG-HRP (Thermo Fisher, 31458, polyclonal, 1:5000)
 WB: Mouse IgG-HRP (Thermo Fisher, 31430, polyclonal, 1:5000)
 WB and IHC: PCK1 pS90 (Signalway, #58006, Rabbit polyclonal, WB: 1:1000, IHC: 1:100)
 WB and IHC: Insig1/2 pS207/S151 (Signalway, #58005, Rabbit polyclonal, WB: 1:1000, IHC: 1:100)
 IP: AKT (Cell signaling, #2920, 40D4, Mouse Monoclonal, IP: 1:100)
 IP: Normal mouse IgG (Santa Cruz Biotechnology, sc-2025, IP: 0.5-4.0 µg, control)
 IP: Normal rabbit IgG (Santa Cruz Biotechnology, sc-2027, IP: 0.5-4.0 µg, control)
 IF: Golgi 97 (Thermo Fisher, A-21270, CDF4, Mouse Monoclonal, IF: 1:200)
 IF and IHC: SREBP1 antibody (Novus, NB100-2215, 2A4, IF: 1:200; IHC: 1:200)
 IF: Rabbit IgG-Alexa Fleor 488 (Invitrogen, A11008, 34732A, Polyclonal, 1:500)
 IF: Rabbit IgG-Alexa Fleor 594 (Invitrogen, A11012, 1810936, Polyclonal, 1:500)
 IF: Mouse IgG-Alexa Fleor 488 (Invitrogen, A11029, 673781, Polyclonal, 1:500)
 IF: Mouse IgG-Alexa Fleor 594 (Invitrogen, A11005, 610868, Polyclonal, 1:500)
 IHC: Ki67 (Sigma-Aldrich, #AB9260, Rabbit polyclonal, IHC: 1:200)

Validation

Antibodies were only used for the application as indicated and organisms verified by the manufactures. Besides, antibodies against PCK1 S90 phosphorylation and Insig1/2 S207/S151 phosphorylation were also double-validated for immunoblotting and IHC analyses by competitive blocking assay with PCK1 pS90 (Extended Data Fig. 2h) and Insig1/2 pS207/S151 peptides (Extended Data Fig. 3j-k) .

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

Hep3B, Huh7, H1993, CHL-1, SNU-398, SNU-475, HL7702, THLE-2 and 293T cells were from ATCC. The U251 and U87 GBM cells used in the experiments were authenticated using short tandem repeat profiling at The University of Texas MD

	Anderson Cancer Center.
Authentication	Cells were authenticated using short tandem repeat profiling at The University of Texas MD Anderson Cancer Center (Houston, Texas)
Mycoplasma contamination	All cells lines are confirmed without Mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified lines were used

Palaeontology

Specimen provenance	<i>Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).</i>
Specimen deposition	<i>Indicate where the specimens have been deposited to permit free access by other researchers.</i>
Dating methods	<i>If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.</i>

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Animals and other organisms

Policy information about [studies involving animals](#); [ARRIVE guidelines](#) recommended for reporting animal research

Laboratory animals	Female, 4-week-old BALB/c athymic nude mice and Wild-type FVB/N mice were used. The use of the animals was approved by the Institutional Review Board at MD Anderson Cancer Center and the Institutional Animal Care and Use Committee (IACUC) of Zhejiang University.
Wild animals	This study did not involve wild animals
Field-collected samples	No sample was collected from the field.

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	Age and Gender are randomly distributed in different groups. Sample numbers are large enough for statistical analysis. Human HCC and adjacent matched nontumor tissue samples from 90 patients (EHBH cohort) were obtained from Eastern Hepatobiliary Surgery Hospital in Shanghai, China. The use of human HCC samples and the relevant database was approved by the Eastern Hepatobiliary Surgery Hospital Research Ethics Committee and complied with all relevant ethical regulations. All tissue samples were collected in compliance with informed consent policy. A total of 90 patients who undertook surgical resection in Eastern Hepatobiliary Surgery Hospital from 2009 to 2011 were included in this study. 69 males and 21 females, with an average (\pm SD) age of 46.7 (\pm 10.3) years undertook surgical resection. All patients had received standard therapies after surgery. Sections of paraffin-embedded human HCC samples were stained with antibodies against AKT pS473, PCK1 pS90, Insig1 pS207/Insig2 pS151, SREBP1, or nonspecific IgG as a negative control. The staining of the tissue sections was quantitatively scored according to the percentage of positive cells and staining intensity. Scores were compared with overall survival duration, defined as the time from date of diagnosis to that of death or last known follow-up examination.
Recruitment	participants were not recruited.

ChIP-seq

Data deposition

- ☐ Confirm that both raw and final processed data have been deposited in a public database such as [GEO](#).
- ☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links <i>May remain private before publication.</i>	<i>For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.</i>
Files in database submission	<i>Provide a list of all files available in the database submission.</i>
Genome browser session (e.g. UCSC)	<i>Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.</i>

Methodology

Replicates	Describe the experimental replicates, specifying number, type and replicate agreement.
Sequencing depth	Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.
Antibodies	Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.
Peak calling parameters	Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.
Data quality	Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.
Software	Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.

Flow Cytometry

Plots

Confirm that:

- ☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).
- ☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).
- ☐ All plots are contour plots with outliers or pseudocolor plots.
- ☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.
Instrument	Identify the instrument used for data collection, specifying make and model number.
Software	Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.
Cell population abundance	Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.
Gating strategy	Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.
<input type="checkbox"/> Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.	

Magnetic resonance imaging

Experimental design

Design type	Indicate task or resting state; event-related or block design.
Design specifications	Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.
Behavioral performance measures	State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).

Acquisition

Imaging type(s)	Specify: functional, structural, diffusion, perfusion.
Field strength	Specify in Tesla
Sequence & imaging parameters	Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.
Area of acquisition	State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	<i>Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).</i>
Normalization	<i>If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.</i>
Normalization template	<i>Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.</i>
Noise and artifact removal	<i>Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).</i>
Volume censoring	<i>Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.</i>

Statistical modeling & inference

Model type and settings	<i>Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).</i>
Effect(s) tested	<i>Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.</i>
Specify type of analysis: <input type="checkbox"/> Whole brain <input type="checkbox"/> ROI-based <input type="checkbox"/> Both	
Statistic type for inference (See Eklund et al. 2016)	<i>Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.</i>
Correction	<i>Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).</i>

Models & analysis

n/a	Involved in the study
<input type="checkbox"/>	<input type="checkbox"/> Functional and/or effective connectivity
<input type="checkbox"/>	<input type="checkbox"/> Graph analysis
<input type="checkbox"/>	<input type="checkbox"/> Multivariate modeling or predictive analysis
Functional and/or effective connectivity	<i>Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).</i>
Graph analysis	<i>Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).</i>
Multivariate modeling and predictive analysis	<i>Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.</i>

Securin-independent regulation of separase by checkpoint-induced shugoshin–MAD2

<https://doi.org/10.1038/s41586-020-2182-3>
Susanne Hellmuth¹, Laura Gómez-H², Alberto M. Pendás² & Olaf Stemmann¹✉

Received: 22 March 2019

Accepted: 31 January 2020

Published online: 8 April 2020

 Check for updates

Separation of eukaryotic sister chromatids during the cell cycle is timed by the spindle assembly checkpoint (SAC) and ultimately triggered when separase cleaves cohesion-mediating cohesin^{1–3}. Silencing of the SAC during metaphase activates the ubiquitin ligase APC/C (anaphase-promoting complex, also known as the cyclosome) and results in the proteasomal destruction of the separase inhibitor securin¹. In the absence of securin, mammalian chromosomes still segregate on schedule, but it is unclear how separase is regulated under these conditions^{4,5}. Here we show that human shugoshin 2 (SGO2), an essential protector of meiotic cohesin with unknown functions in the soma^{6,7}, is turned into a separase inhibitor upon association with SAC-activated MAD2. SGO2–MAD2 can functionally replace securin and sequesters most separase in securin-knockout cells. Acute loss of securin and SGO2, but not of either protein individually, resulted in separase deregulation associated with premature cohesin cleavage and cytotoxicity. Similar to securin^{8,9}, SGO2 is a competitive inhibitor that uses a pseudo-substrate sequence to block the active site of separase. APC/C-dependent ubiquitylation and action of the AAA-ATPase TRIP13 in conjunction with the MAD2-specific adaptor p31^{comet} liberate separase from SGO2–MAD2 in vitro. The latter mechanism facilitates a considerable degree of sister chromatid separation in securin-knockout cells that lack APC/C activity. Thus, our results identify an unexpected function of SGO2 in mitotically dividing cells and a mechanism of separase regulation that is independent of securin but still supervised by the SAC.

In all eukaryotic cells, anaphase is triggered when chromosomal cohesin is cleaved by the essential Cys-endopeptidase separase^{3,10}. To prevent the premature loss of sister chromatid cohesion, separase needs to be tightly controlled. Separase is competitively inhibited by association with securin for most of the cell cycle. Only in metaphase does the E3 anaphase-promoting complex or cyclosome (APC/C) mediate the degradation of securin via the ubiquitin-proteasome system, thereby activating separase¹¹. The destruction of securin is timed by the SAC, which keeps the APC/C co-activator CDC20 inactive until all kinetochores are properly attached to spindle microtubules¹.

SGO2 is a prominent interactor of separase

Unexpectedly, securin is not essential in human cells or mice^{4,5}. This can partially be explained by CDK1–cyclin B1-dependent regulation of separase^{12–17}. Mouse *Cdc20*^{−/−} embryos arrest in metaphase with cohered chromosomes because they cannot degrade either cyclin B1 or securin¹⁸. Notably, double knockout of *Cdc20* and *Pttg1* (which encodes securin) resulted in arrest with separated sister chromatids; this defect was rescued by constitutive activation of the SAC¹⁸, which suggest that there is a SAC-dependent but securin-independent mechanism to control separase. Rather than being stimulated by the SAC, the binding of CDK1–cyclin B1 to separase is dampened by

phosphorylation of cyclin B during early mitosis¹⁹. However, a link between the SAC and the cohesin protector shugoshin (SGO) had previously been identified in that human SGO2—similar to CDC20—is bound by the essential SAC component MAD2⁷. Notably, mouse SGO2 and separase interacted when co-expressed in Hek293 cells (Extended Data Fig. 1a; see, however, Extended Data Fig. 1b). These findings led to the idea that SAC-activated MAD2 could enable SGO2 to bind and inhibit separase. Indeed, when endogenous human separase was isolated by immunoprecipitation (IP) from untransfected, prometaphase-arrested Hek293T or untransformed RPE-1 cells, SGO2 (but not the related SGO1) and MAD2 co-purified, along with the known interactors securin and cyclin B1 (Fig. 1a, Extended Data Fig. 1c). In contrast to separase, SGO2 and MAD2 were undetectable in a securin IP and, vice versa, securin and cyclin B1 were absent from an SGO2 IP (Fig. 1a). With previous results²⁰, these data argue that three mutually exclusive complexes co-exist in human mitotic cells: separase–securin, separase–CDK1–cyclin B1 and separase–SGO2–MAD2. Following immunodepletion of the three inhibitors from taxol-arrested Hek293T, HCT116 and HeLa-K cells, we quantified the relative amounts of associated separase. On average, 59, 35, and 6% of total separase was sequestered by securin, SGO2, and cyclin B1, respectively (Extended Data Fig. 1d). Notably, in mitotic *PTTG1*^{−/−} cells, most separase (85%) was in complex with SGO2 (Fig. 1b).

¹Chair of Genetics, University of Bayreuth, Bayreuth, Germany. ²Molecular Mechanisms Program, Centro de Investigación del Cáncer and Instituto de Biología Molecular y Celular del Cáncer (CSIC-Universidad de Salamanca), Salamanca, Spain. ✉e-mail: olaf.stemmann@uni-bayreuth.de

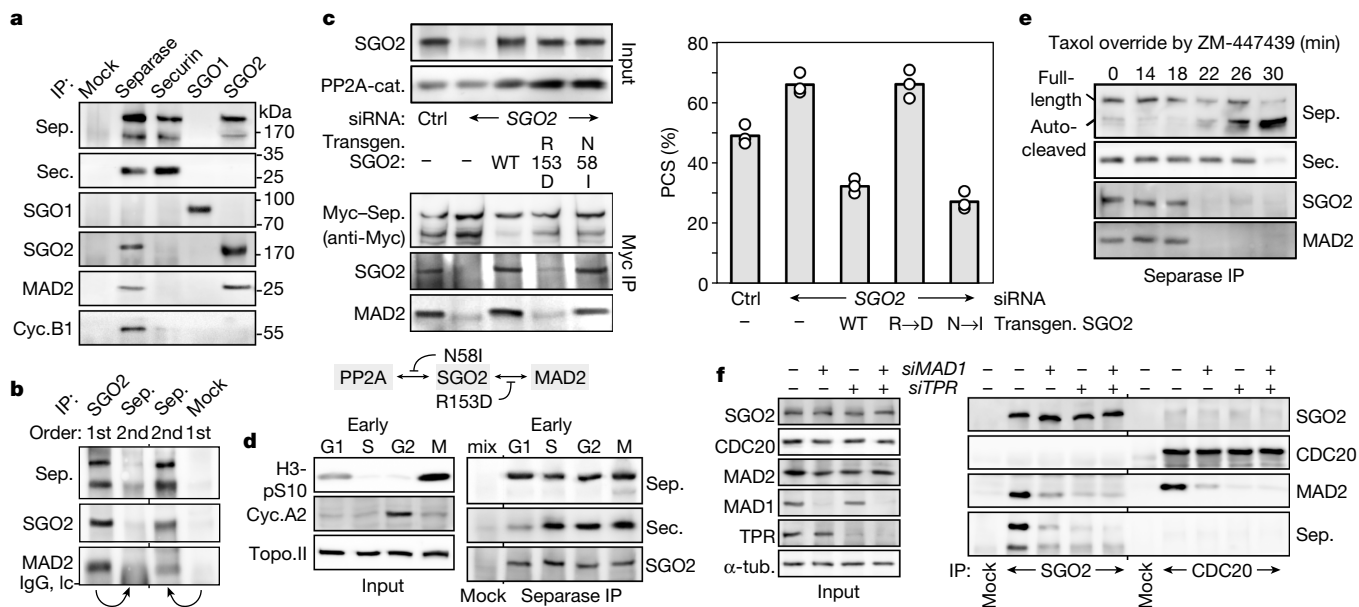


Fig. 1 | MAD2-dependent binding of human SGO2 to separase. **a**, Endogenous proteins were immunoprecipitated (IP) from taxol-arrested Hek293T cells and analysed by immunoblotting. Mock, nonspecific IgG. Sep., separase; sec., securin; cyc.B1, cyclin B1. **b**, Lysates from taxol-arrested *PTTG1*^{-/-} cells were subjected to consecutive immunodepletions and analysed by immunoblotting. **c**, Transgenic Hek293 cells depleted of SGO2 or control-treated (Ctrl) were transfected to express siRNA-resistant SGO2 variants as indicated (transgen. SGO2: wild-type (WT), R153D or N581I), induced with doxycycline to express CDK1–cyclin B1-resistant, stabilized Myc–separase (S1126A), arrested in

prometaphase, and then analysed by IP–immunoblotting (left) and chromosome spreading (right). Bars show mean of three independent experiments (dots). Cat., catalytic subunit. **d**, HeLa-K cells synchronized in the indicated cell cycle phases were subjected to IP–immunoblotting. **e**, Taxol-arrested HeLa-K cells were released with ZM-447439 and subjected to time-resolved IP–immunoblotting. **f**, Control, MAD1-depleted, or TPR-depleted, thymidine-arrested HeLa-K cells were analysed by IP–immunoblotting. α-tub., α-tubulin.

MAD2 enables SGO to bind separase

To test whether binding of SGO2 to separase required MAD2, we used RNA interference (RNAi) to deplete cells of endogenous SGO2 and replaced it with small interfering RNA (siRNA)-resistant, transgene-encoded variants. Subsequent immunoprecipitation of co-expressed Myc–separase showed that the MAD2-binding-deficient SGO2(R153D) was unable to interact with separase, whereas SGO2(N581I), which cannot interact with protein phosphatase 2A (PP2A)⁷, still bound to MAD2 and separase (Fig. 1c). The SAC–shugoshin link is conserved in *Xenopus*, with the difference that here SGO1 rather than SGO2 binds MAD2 and separase⁷ (Extended Data Fig. 2). Thus, the separase–shugoshin interaction depends on active MAD2 and is conserved in vertebrates.

The separase–SGO2–MAD2 complex was present in cells arrested in G1, early S, G2, and prometaphase (Fig. 1d) but not in taxol-treated HeLa-K cells that were driven into an anaphase-like state by SAC abrogation with the aurora B kinase inhibitor ZM-447439 (Fig. 1e). This cell cycle distribution mirrored that of CDC20–MAD2, the formation of which in interphase requires MAD1-dependent MAD2 activation at nuclear pore complexes (NPCs)²¹. Consistently, separase–SGO2–MAD2, similar to CDC20–MAD2, became (almost) undetectable in unsynchronized HeLa-K cells depleted of MAD1 and/or the NPC component TPR (Fig. 1f). Thus, conformationally activated MAD2 enables SGO2 to sequester separase through all of the cell cycle except for a short period of SAC inactivity during late mitosis.

Separase deregulation upon loss of securin and SGO

Overexpression of a CDK1–cyclin B1-resistant and stabilized separase (S1126A) variant causes premature sister chromatid separation (PCS) in Hek293T cells^{12,22}. This PCS phenotype was aggravated by siRNA-mediated depletion of endogenous SGO2 and alleviated

by simultaneous slight overexpression of wild-type SGO2 from an siRNA-resistant transgene (Fig. 1c). SGO2(R153D) did not rescue PCS in this cellular assay, whereas transgenic SGO2(N581I) remained functional. Together, these findings suggest that SGO2 might indeed have an inhibitory effect on separase and that this effect requires binding of MAD2 but not PP2A to SGO2.

By recruiting PP2A to (peri)centromeres, SGO2 exerts essential cohesin protective functions throughout meiosis I but, similar to securin, it is dispensable in somatic cells^{6,7}. If securin and SGO2–MAD2 could functionally replace each other as crucial negative regulators of separase, then co-depletion of securin and SGO2 should result in detrimental deregulation of separase. First, we assessed overall effects on cell viability and proliferation using clonogenic assays. As expected, knockdown of securin or SGO2 alone had no effect or only a small inhibitory effect, respectively, on HeLa-K colony formation (Fig. 2a, Extended Data Fig. 3a). By sharp contrast, hardly any clones grew when both separase interactors were depleted at the same time. Similarly, depletion of SGO2 virtually extinguished colony formation in *PTTG1*^{-/-} cells, whereas it only halved colony numbers in the parental HCT116 cell line (Extended Data Fig. 3b). The same tendencies were found for Hek293T cells, although the effects were less pronounced (Extended Data Fig. 3c). Live imaging of histone H2B–eGFP-expressing HeLa-K cells revealed that the individual knockdowns did not affect mitosis; however, when both securin and SGO2 were missing, the ability to form proper metaphase plates was markedly compromised (Extended Data Fig. 4a). HeLa-K cells lacking securin and SGO2 were also marked by the otherwise uncommon absence of cohesin from early mitotic chromatin (Extended Data Fig. 4b). Individual knockdown of securin or SGO2 had no effect on cohesin, whereas simultaneous removal of both resulted in PCS as judged by chromosome spreads from prometaphase-arrested Hek293T cells (Extended Data Fig. 4c, d). The PCS phenotype was further fortified by chemical inhibition (using epigallocatechin-3-gallate (EGCG)) of PIN1, a peptidyl–prolyl–isomerase required for

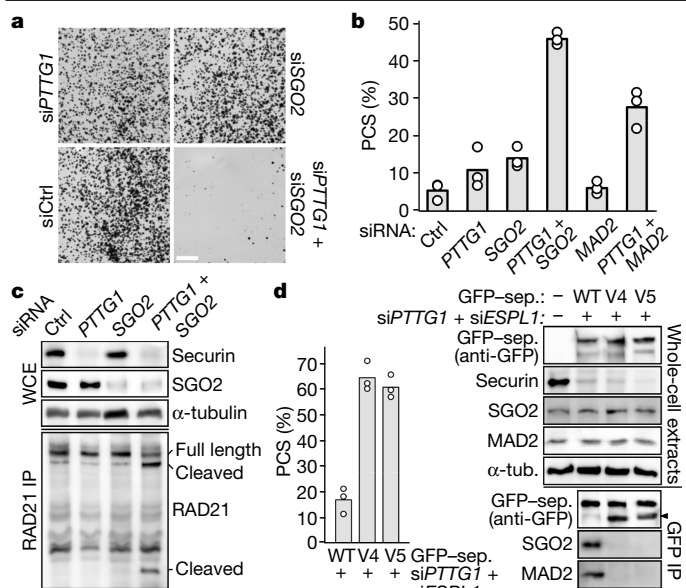


Fig. 2 | Co-depletion of securin and SGO2 is cytotoxic and results in premature sister chromatid separation and cohesin cleavage. **a**, Exemplary photographs of clonogenic assay with HeLa-K cells transfected with the indicated siRNAs. Scale bar, 1 cm. **b**, siRNA-transfected Hek293T cells were supplemented with taxol and EGCG and analysed by chromosome spreading. Bars show mean of three independent experiments (dots). **c**, siRNA-transfected HeLa-K cells were taxol-treated and analysed by IP-immunoblotting. WCE, whole-cell extracts. **d**, Securin-depleted (siPTTG1) and separase-depleted (siESPL1) Hek293T cells expressing siRNA-resistant, transgenic separase (wild-type, ΔEEL (V4), or MxxIxEE to AxxAxA (V5)) were treated with taxol and EGCG and analysed by chromosome spreading (left) and IP-immunoblotting (right). Bars show mean of three independent experiments (dots). Arrowhead marks auto-cleaved separase.

CDK1–cyclin B1-dependent inhibition of separase¹⁹ (Fig. 2b, Extended Data Fig. 4c, d). In keeping with the separase–SGO2 interaction being dependent on MAD2, partial removal of MAD2, which alone had no effect, inhibited colony formation and exacerbated PCS in conjunction with RNAi of securin (Fig. 2b, Extended Data Figs. 3, 4c, d). All of these phenotypes could be explained by precocious separase-dependent cleavage of cohesin. Indeed, consistent with a previous study in yeast²³, the characteristic fragments of the RAD21 subunit of cohesin were detected in prometaphase-arrested cells if—and only if—both securin and SGO2 had been depleted (Fig. 2c). Likewise, a transiently expressed separase activity sensor was already maximally cleaved in securin- and SGO2-depleted HeLa-K cells in a taxol arrest, whereas in mock-depleted cells the sensor was cleaved only upon addition of ZM-447439 (Extended Data Fig. 4e). Screening of separase mutants identified two variants (V4 and V5) that could not interact with SGO2–MAD2 (Extended Data Fig. 5). When endogenous separase was replaced by these variants, depletion of securin was sufficient to induce PCS (Fig. 2d). This indicates that a direct separase inhibitory function of SGO2 works redundantly with securin to prevent PCS.

In interphase, separase is excluded from the nucleus but has an established function in centriole disengagement²⁴. Therefore, we assessed this licensing step of centrosome duplication rather than cleavage of chromosomal cohesin upon depletion of securin and SGO2. Premature centriole disengagement in G2-phase was increased threefold in Hek293T cells lacking securin and SGO2 relative to singly- or mock-depleted Hek293T cells (45% versus 10–15%, respectively) (Extended Data Fig. 6). Thus, cells lacking securin and SGO2 already contain active separase in interphase and, hence, are expected to lose sister chromatid cohesion immediately upon breakdown of the nuclear envelope.

Shugoshins are characterized not only by an N-terminal coiled-coil domain but also by a C-terminal SGO-box that binds to BUB1-phosphorylated histone H2A²⁵. Phosphorylated, but not unphosphorylated, H2A peptide bound immobilized SGO2 (Extended Data Fig. 7a–c). Notably, pre-incubation with phospho-H2A but not with unphosphorylated H2A or a chemically similar but irrelevant phospho-H3 peptide suppressed the ability of SGO2 to bind MAD2 (Extended Data Fig. 7a–d). Conversely, pre-incubation of immobilized SGO2 with MAD2 prevented subsequent interaction between phospho-H2A and SGO2 (Extended Data Fig. 7e, f). Phospho-H2A did bind to SGO2 when previous SGO2–MAD2 complex formation was not possible owing to the expression of SGO2(R153D) or a variant of MAD2 lacking the C-terminal domain (MAD2ΔC) instead of the corresponding wild-type proteins. Thus, MAD2 and phospho-H2A bind SGO2 in a mutually exclusive manner. We propose that different pools of SGO2 either associate with chromatin or bind MAD2 and inhibit separase.

SGO is a pseudosubstrate inhibitor of separase

Incubation of the human separase–securin complex in securin-degrading, anaphase-like *Xenopus laevis* egg extracts followed by affinity purification of separase generates active protease that specifically cleaves ³⁵S-labelled RAD21¹⁷. Pre-incubation of this separase with in vitro-expressed SGO2 and MAD2 purified from *Escherichia coli* blocked cleavage of RAD21 (Fig. 3a). The same effect was seen when SGO2(N53I) was used instead of wild-type SGO2, but not when SGO2 or MAD2 was omitted or replaced by SGO2(R153D) or MAD2ΔC, respectively. Thus, the PP2A-independent but MAD2-dependent inhibition of human separase by SGO2 can be recapitulated in vitro.

We investigated whether shugoshin inhibits separase in the same way as securin—by occupying the catalytic site with a non-cleavable pseudo-substrate sequence^{8,9}. Using interaction-blocking antibodies, protein fragments and point mutations, we mapped sites within *X. laevis* Sgo1 that are important for separase interaction (Extended Data Fig. 8). Nearby putative pseudosubstrate sites (ϕExxX, with ϕ denoting a hydrophobic residue, x denoting any residue and X denoting any residue except R) were then changed into consensus sites (ϕExxR) and the resulting variants screened for cleavage by active *X. laevis* separase. Notably, *X. laevis* Sgo1(S135R) (but not Sgo1(F288R) or the wild type) was cleaved by separase, and this cleavage was much more pronounced in the presence of wild-type MAD2 than in the presence of MAD2ΔC (Extended Data Fig. 9a).

To confirm this finding independently, we switched back to the human system and tested whether a specific ϕExxX-to-ϕExxR mutation could also turn human SGO2 into a separase substrate. Indeed, ³⁵S-labelled SGO2(M114R) (but not SGO2(F95R) or SGO2(S126R)) was fragmented in the presence of active separase and wild-type MAD2 (Fig. 3b). Cleavage of SGO2(M114R)—similar to that of RAD21²⁶—was further enhanced upon phosphorylation by Polo-like kinase 1. To test in vivo cleavage, we transfected HeLa-K cells to express Flag-tagged variants of SGO2 or securin. Subsequent immunoprecipitation from prometaphase lysates demonstrated that both securin(F118R) (positive control) and SGO2(M114R) were fragmented, whereas the wild-type proteins remained unprocessed (Fig. 3c). Cleavage of SGO2(M114R) was separase-specific because it was suppressed by introduction of a second mutation that compromised the interaction of SGO2 with separase (amino acids 239–242 of SGO2 to Ala; Fig. 3c, Extended Data Fig. 9b). Thus, shugoshin resembles securin in acting as a competitive inhibitor of separase but differs in that it requires MAD2 binding to do so.

X. laevis Sgo1 and human SGO2 are very different in sequence (21% similarity) and length (663 versus 1,265 residues). However, the relative order of and distance between functional elements are the same in both proteins (Extended Data Fig. 9c). This suggests that MAD2 binding and separase inhibition evolved before duplication of a primordial SGO gene and were later lost randomly from one SGO gene but retained in the other owing to selective pressure.

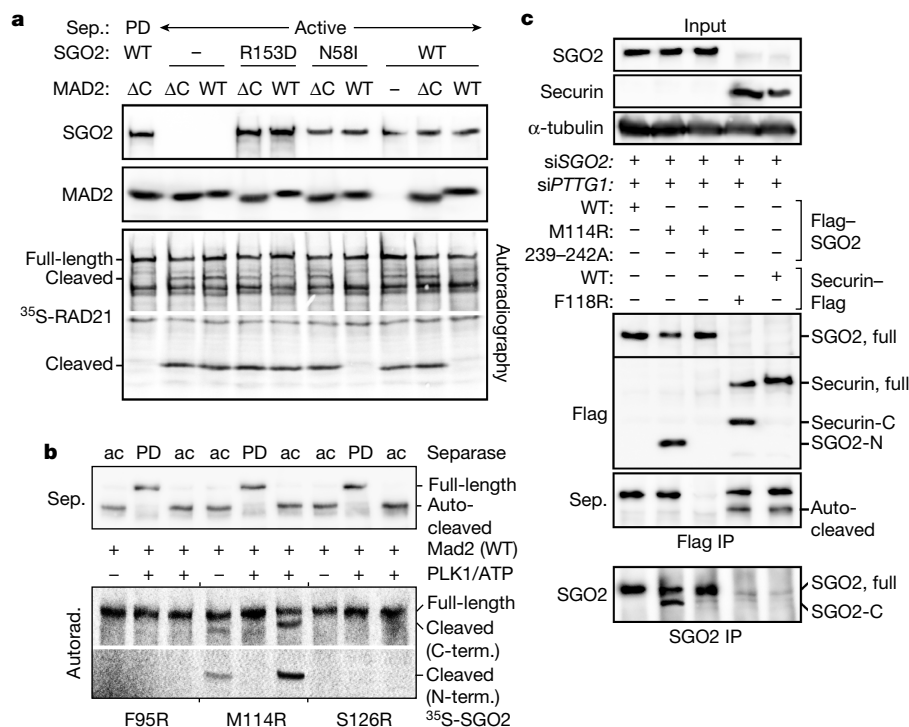


Fig. 3 | SGO2 is a MAD2-dependent, competitive inhibitor of separase. a, Protease-dead (PD) or active separase was incubated with variants of recombinant SGO2 and MAD2 and assayed for its ability to cleave ^{35}S -labelled RAD21. Relevant upper and lower parts of the same gel are shown. **b**, ^{35}S -labelled SGO2 variants were incubated with MAD2 and, where indicated, Polo-like kinase 1 (PLK1) before being assayed for in vitro cleavage by active (Ac)

separase. Relevant upper and lower parts of the same gel are shown. **c**, SGO2- and securin-depleted HeLa-K cells were transfected to express the indicated siRNA-resistant, Flag-tagged SGO2 or securin variants, taxol-arrested and analysed by IP-immunoblotting. Separase-induced N- and/or C-terminal cleavage fragments of SGO2 and securin are labelled -N and -C.

TRIP13 liberates separase from SGO2-MAD2

In contrast to securin, human SGO2 is not (or only slowly and incompletely) degraded in late mitosis. This raises the question of how separase is liberated from SGO2-MAD2 when cells are ready to undergo anaphase. The CDC20- and MAD2-containing mitotic checkpoint complex (MCC) is disassembled by the combined action of the AAA-ATPase TRIP13 and its MAD2-specific adaptor p31^{comet} (ref. 27). We tested whether this molecular machine could also dismantle the separase-SGO2-MAD2 complex. The complex was immunoprecipitated from securin-depleted, taxol-arrested HeLa-K cells using antibodies against separase and incubated with different combinations of recombinant TRIP13 and p31^{comet} variants. Beads were then washed to remove detached proteins, incubated with ^{35}S -RAD21 to assay for activity of the immobilized separase, and finally analysed for retained proteins. Notably, wild-type TRIP13 and p31^{comet} quantitatively displaced SGO2 and MAD2 from separase, thereby leaving it proteolytically active (Fig. 4a). Whereas TRIP13 alone partially disassembled separase-SGO2-MAD2, the Walker-A and -B mutant variants of TRIP13 were inactive even in the presence of p31^{comet}. Activation of separase by TRIP13 was also prevented by p31^{comet} variants that were defective in MAD2 or TRIP13 interaction. To assess the role of TRIP13 and p31^{comet} in vivo, we additionally transfected securin-depleted HeLa-K cells with siRNAs against TRIP13 and p31^{comet}, synchronized the cells in prometaphase with taxol and then released them using ZM-447439. Unexpectedly, late mitotic events, such as the de-phosphorylation of CDC27 and histone H3 and the degradation of cyclin B1, were only slightly delayed or occurred largely on schedule in cells lacking TRIP13 and p31^{comet} relative to control-treated cells (Fig. 4b, top). By contrast, the separase-SGO2-MAD2 complex was markedly stabilized in the absence of TRIP13 and p31^{comet}, as revealed by separase immunoprecipitation

and immunoblotting (Fig. 4b, bottom). Thus, separase-SGO2-MAD2 is actively dismantled by TRIP13 and p31^{comet} and might depend on this molecular machine for its disassembly in late mitosis even more than the MCC.

Sister separation without APC/C activity

The above results suggest that sister chromatid separation in the absence of APC/C^{CDC20} activity could be possible when separase is chiefly controlled by SGO2-MAD2 instead of securin. To test this prediction, we supplemented taxol-arrested *PTTG1*^{-/-} and parental HCT116 cells with the two APC/C inhibitors proTame and Apcin (or carrier solvent), released the cells by adding ZM-447439, and analysed them by time-resolved immunoprecipitation-immunoblotting and chromosome spreading. Thirty-five minutes after inhibition of aurora B kinase, up to 60% of chromosomes were separated in *PTTG1*^{-/-} cells but only 20% on average in parental HCT116 cells, despite the persistence of CDC27 phosphorylation and cyclin B1 in both (Fig. 4c, d). Whereas levels of separase-associated securin also stayed constant in APC/C-inhibited HCT116 cells, SGO2 and MAD2 (which appeared to be overexpressed in *PTTG1*^{-/-} cells) disappeared from separase irrespective of securin status (Fig. 4d). Consistent with SGO2-MAD2 being the primary inhibitor of separase, auto-cleavage of separase was strongly increased in *PTTG1*^{-/-} cells. Thus, activation of SGO2-MAD2-inhibited separase occurs at least partially independently of APC/C. However, sister chromatid separation was more effective in the absence of proTame and Apcin, even in *PTTG1*^{-/-} cells (Fig. 4c). Given that APC/C-associated MCC is disassembled also upon ubiquitylation of CDC20²⁸, and given the recent identification of separase and SGO2 as APC/C interactors or substrates²⁹, we tested whether separase-SGO2-MAD2

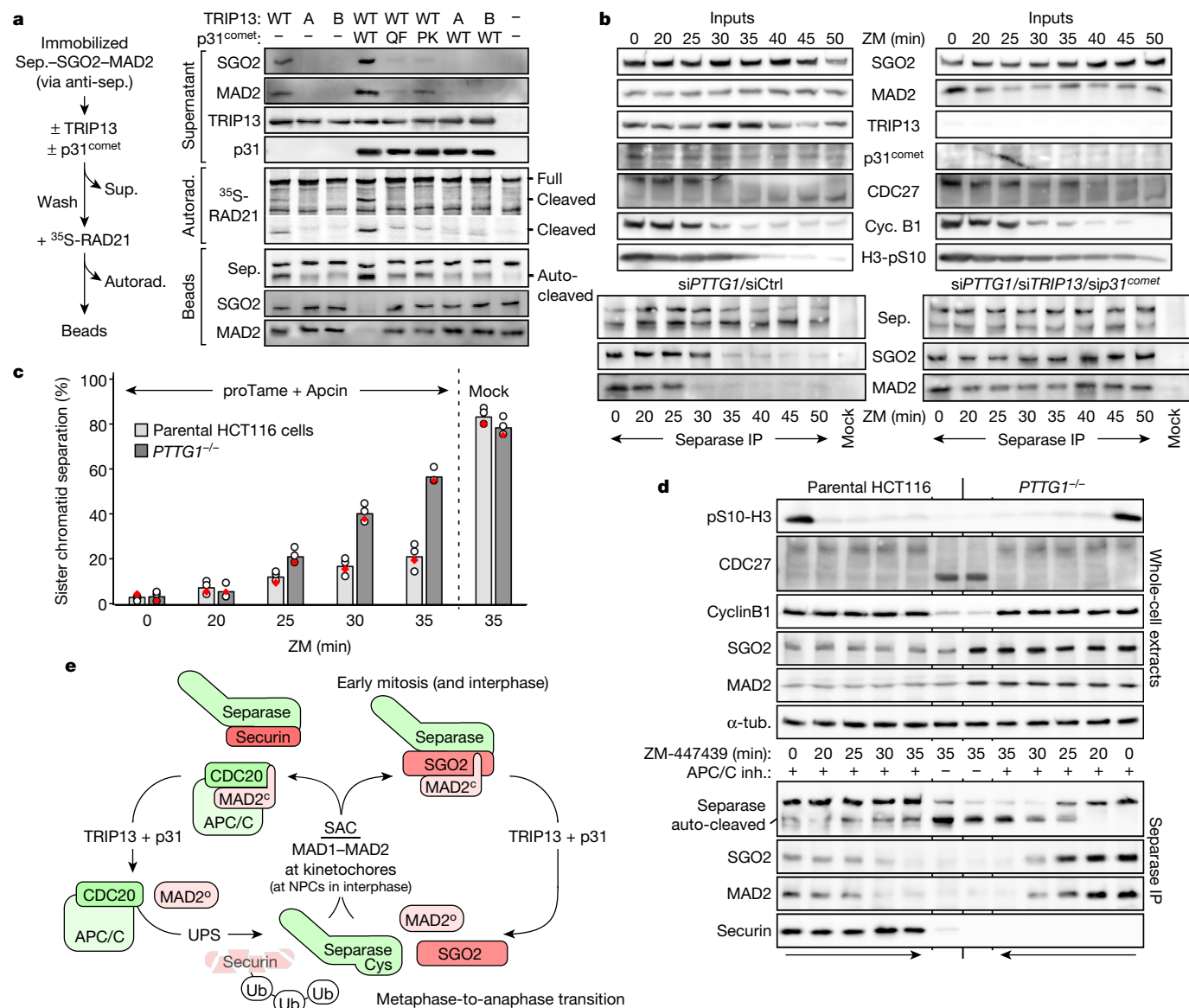


Fig. 4 | TRIP13–p31^{comet}-dependent disassembly liberates separase from SGO2–MAD2. **a**, Experimental scheme (left) and corresponding immunoblots and autoradiograph (right; relevant upper and lower parts of the same gel are shown). **A**, TRIP13(G184A); **B**, TRIP13(E253Q); **QF**, p31^{comet}(Q83A, F191A); **PK**, p31^{comet}(P228A, K229A). **b**, siRNA-transfected, taxol-arrested HeLa-K cells were released using ZM-447439 (ZM) and subjected to time-resolved IP–immunoblotting. **c**, **d**, Taxol-arrested PTTG1^{-/-} and parental HCT116 cells were supplemented with ZM and APC/C inhibitors or carrier solvent (mock) and analysed by chromosome spreading (**c**) and time-resolved IP–immunoblotting (**d**). Bars show mean of four independent experiments (dots). Red diamonds

indicate experiment shown in **d**. **e**, Model of bifurcated regulation of separase. Activation of MAD2 by SAC signalling inhibits APC/C^{CDC20}, thereby stabilizing securin, and enables shugoshin to directly inhibit separase. Liberation of separase from securin requires TRIP13–p31^{comet}-dependent dissociation of MAD2 from CDC20 followed by APC/C-dependent degradation of securin. By contrast, TRIP13–p31^{comet}-dependent dissociation of MAD2 from shugoshin leads to direct activation of associated separase. The alternative dissociation of MAD2 from its targets by APC/C^{CDC20}-dependent ubiquitylation and the CDK1–cyclin B1-dependent inhibition of separase are omitted for clarity.

could also be dismantled by APC/C-dependent ubiquitylation in vitro. Indeed, incubation of the immobilized complex with E1, E2s (UBE2C and UBE2S) and active APC/C^{CDC20} in the presence of ubiquitin and ATP displaced SGO2 and MAD2 and rendered separase proteolytically active (Extended Data Fig. 10). Dissociation was accompanied by ubiquitylation of separase. The complex stayed intact, however, when UBE2C was replaced by a dominant-negative variant or when APC/C^{CDC20} was omitted. These results suggest that APC/C-dependent ubiquitylation represents a second mode of separase–SGO2–MAD2 disassembly.

We propose a model in which there is bifurcated regulation of separase downstream of the SAC (Fig. 4e). Next to the APC/C^{CDC20}–securin

axis, a second major branch is represented by mammalian SGO2 (Sgo1 in amphibians) which is turned into a direct, competitive inhibitor of separase by SAC-activated MAD2. Both branches use TRIP13–p31^{comet} (and APC/C-dependent ubiquitylation) for disassembly of their respective MAD2-containing complexes. However, while this liberates separase from shugoshin, the canonical branch additionally requires proteasomal destruction of securin. At least in the cell lines tested here, securin and SGO2 can each compensate for loss of the other. The reason for this seeming redundancy remains to be clarified, but the different requirements for protein degradation might make it beneficial at times to rely on one or the other mode of separase regulation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2182-3>.

- Musacchio, A. The molecular biology of spindle assembly checkpoint signaling dynamics. *Curr. Biol.* **25**, R1002–R1018 (2015).
- Nasmyth, K. & Haering, C. H. Cohesin: its roles and mechanisms. *Annu. Rev. Genet.* **43**, 525–558 (2009).
- Uhlmann, F., Wernic, D., Poupart, M. A., Koonin, E. V. & Nasmyth, K. Cleavage of cohesin by the CD clan protease separin triggers anaphase in yeast. *Cell* **103**, 375–386 (2000).
- Mei, J., Huang, X. & Zhang, P. Securin is not required for cellular viability, but is required for normal growth of mouse embryonic fibroblasts. *Curr Biol* **11**, 1197–1201 (2001).
- Pfleghaar, K., Heubes, S., Cox, J., Stemmann, O. & Speicher, M. R. Securin is not required for chromosomal stability in human cells. *PLoS Biol.* **3**, e416 (2005).
- Llano, E. et al. Shugoshin-2 is essential for the completion of meiosis but not for mitotic cell division in mice. *Genes Dev.* **22**, 2400–2413 (2008).
- Orth, M. et al. Shugoshin is a Mad1/Cdc20-like interactor of Mad2. *EMBO J.* **30**, 2868–2880 (2011).
- Boland, A. et al. Cryo-EM structure of a metazoan separase-securin complex at near-atomic resolution. *Nat. Struct. Mol. Biol.* **24**, 414–418 (2017).
- Lin, Z., Luo, X. & Yu, H. Structural basis of cohesin cleavage by separase. *Nature* **532**, 131–134 (2016).
- Wirth, K. G. et al. Separase: a universal trigger for sister chromatid disjunction but not chromosome cycle progression. *J. Cell. Biol.* **172**, 847–860 (2006).
- Zou, H., McGarry, T. J., Bernal, T. & Kirschner, M. W. Identification of a vertebrate sister-chromatid separation inhibitor involved in transformation and tumorigenesis. *Science* **285**, 418–422 (1999).
- Boos, D., Kuffer, C., Lenobel, R., Korner, R. & Stemmann, O. Phosphorylation-dependent binding of cyclin B1 to a Cdc6-like domain of human separase. *J. Biol. Chem.* **283**, 816–823 (2008).
- Hellmuth, S. et al. Positive and negative regulation of vertebrate separase by Cdk1-cyclin B1 may explain why securin is dispensable. *J. Biol. Chem.* **290**, 8002–8010 (2015).
- Huang, X. et al. Preimplantation mouse embryos depend on inhibitory phosphorylation of separase to prevent chromosome missegregation. *Mol. Cell. Biol.* **29**, 1498–1505 (2009).
- Huang, X. et al. Inhibitory phosphorylation of separase is essential for genome stability and viability of murine embryonic germ cells. *PLoS Biol.* **6**, e15 (2008).
- Huang, X., Hatcher, R., York, J. P. & Zhang, P. Securin and separase phosphorylation act redundantly to maintain sister chromatid cohesion in mammalian cells. *Mol. Biol. Cell* **16**, 4725–4732 (2005).
- Stemmann, O., Zou, H., Gerber, S. A., Gygi, S. P. & Kirschner, M. W. Dual inhibition of sister chromatid separation at metaphase. *Cell* **107**, 715–726 (2001).
- Li, M., York, J. P. & Zhang, P. Loss of Cdc20 causes a securin-dependent metaphase arrest in two-cell mouse embryos. *Mol. Cell. Biol.* **27**, 3481–3488 (2007).
- Hellmuth, S. et al. Human chromosome segregation involves multi-layered regulation of separase by the peptidyl-prolyl-isomerase Pin1. *Mol. Cell* **58**, 495–506 (2015).
- Gorr, I. H., Boos, D. & Stemmann, O. Mutual inhibition of separase and Cdk1 by two-step complex formation. *Mol. Cell* **19**, 135–141 (2005).
- Rodríguez-Bravo, V. et al. Nuclear pores protect genome integrity by assembling a premitotic and Mad1-dependent anaphase inhibitor. *Cell* **156**, 1017–1031 (2014).
- Holland, A. J. & Taylor, S. S. Cyclin-B1-mediated inhibition of excess separase is required for timely chromosome disjunction. *J. Cell Sci.* **119**, 3325–3336 (2006).
- Clift, D., Bizzari, F. & Marston, A. L. Shugoshin prevents cohesin cleavage by PP2A(Cdc55)-dependent inhibition of separase. *Genes Dev.* **23**, 766–780 (2009).
- Tsou, M. F. & Stearns, T. Mechanism limiting centrosome duplication to once per cell cycle. *Nature* **442**, 947–951 (2006).
- Kawashima, S. A., Yamagishi, Y., Honda, T., Ishiguro, K. & Watanabe, Y. Phosphorylation of H2A by Bub1 prevents chromosomal instability through localizing shugoshin. *Science* **327**, 172–177 (2010).
- Hauf, S. et al. Dissociation of cohesin from chromosome arms and loss of arm cohesion during early mitosis depends on phosphorylation of SA2. *PLoS Biol.* **3**, e69 (2005).
- Eytan, E. et al. Disassembly of mitotic checkpoint complexes by the joint action of the AAA-ATPase TRIP13 and p31^{comet}. *Proc. Natl Acad. Sci. USA* **111**, 12019–12024 (2014).
- Reddy, S. K., Rape, M., Margansky, W. A. & Kirschner, M. W. Ubiquitination by the anaphase-promoting complex drives spindle checkpoint inactivation. *Nature* **446**, 921–925 (2007).
- Bakos, G. et al. An E2-ubiquitin thioester-driven approach to identify substrates modified with ubiquitin and ubiquitin-like molecules. *Nat. Commun.* **9**, 4776 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Antibodies

Antibodies generated within this study are from Charles River Laboratories. Antigenic peptides (Bachem) were coupled via terminal Cys to maleimide-activated KLH (ThermoFisher) before immunization. Antibodies were affinity-purified against immobilized antigens (proteins coupled to NHS-activated sepharose (GE Healthcare) or peptides coupled to Sulfo-link (ThermoFisher)). The following antibodies were used for immunoblotting according to standard protocols. Antibodies directed against human proteins: rabbit or guinea pig anti-SGO2 (1.5 µg/ml; 'ab1'; anti-DVPPRESHSKQSSKC), rabbit anti-SGO2 (1 µg/ml; 'ab2'; anti-KSEDLSSERTSRRRC), guinea pig anti-TRIP13 (raised against full-length TRIP13), rabbit anti-p31^{comet} (raised against isoform 2 of full-length p31^{comet}), rabbit anti-separase¹⁷, mouse anti-securin (1:1,000; MBL, DCS-280), mouse anti-Flag (1:2,000; Sigma-Aldrich, M2), rabbit anti-SGO2 (1:1,000; Bethyl, A301-262A), rabbit anti-SGO1 (1:500, Abcam ab21633), mouse anti-MAD2 (1:800; Santa Cruz Biotechnology, 17D10), rabbit anti-MAD2 (1:1,000; Bethyl, A300-300A), mouse anti-MAD1 (1:1,000; Sigma-Aldrich, 9B10), mouse anti-APC7 (1:800; ThermoFisher, PA5-20948), rabbit anti-phosphoSer10-histone H3 (H3-pS10; 1:1,000; Millipore, 06-570), mouse anti-PP2A-C (1:1,000; Millipore, 1D6), mouse anti-cyclin B1 (1:1,000; Millipore, 05-373), goat anti-CDC27³⁰, mouse anti-topoisomerase IIα (1:1,000; Enzo Life Sciences, 1C5), mouse anti-cyclin A2 (1:200; Santa Cruz Biotechnology, 46B11), mouse anti-RAD21 (1:800; Santa Cruz Biotechnology, B-2), rabbit anti-RAD21 (1:1,000; Bethyl, A300-080A). Antibodies directed against *Xenopus* proteins: rabbit anti-Mad2 (raised against full-length protein), four different rabbit anti-Sgo1 (raised against amino acids 200–300, 300–400, 400–500, and 500–600 of *X. laevis* Sgo1)⁷, rabbit anti-separase³¹. Other antibodies: mouse anti-Myc (hybridoma supernatant 1:50; DSHB, 9E10), rat anti-HA (1:2,000; Roche, 3F10), rabbit anti-ovalbumin (1:1,000; ThermoFisher, PA1-196), mouse anti-ubiquitinated proteins (1:1,000; Millipore, FK2), mouse anti-GFP³², and mouse anti-α-tubulin (hybridoma supernatant 1:200; DSHB, 12G10). For immunoprecipitation experiments, the following affinity matrices and antibodies were used: mouse anti-Myc agarose (Sigma-Aldrich, 4A6), mouse anti-Flag M2-agarose (Sigma-Aldrich), mouse anti-RAD21 coupled to protein G sepharose (GE Healthcare). Rabbit anti-separase (human), rabbit anti-securin³⁰, rabbit anti-SGO2 and rabbit anti-SGO1 (human) were coupled to protein A sepharose (GE healthcare). To precipitate *X. laevis* separase from CSF (cytostatic factor) extract, the corresponding separase antibody or nonspecific rabbit IgG was coupled to magnetic protein A Dynabeads (Invitrogen). For non-covalent coupling of antibodies to beads, 10 µl of the respective matrix was rotated with 2–5 µg antibody for 90 min at room temperature and then washed three times with LP2 lysis buffer. For immunofluorescence microscopy (IFM), rabbit anti-Cap-E (anti-CAKSKAKPPKGAHVEV) and mouse anti-RAD21 (1:500; Millipore, 05-908) were used. Isolated centrosomes were stained with rabbit anti-centrin-2, guinea-pig anti-C-Nap1 and mouse anti-γ-tubulin (Sigma-Aldrich, GTU-88) as previously described³³. Secondary antibodies (all 1:500): Cy3 donkey anti-guinea pig IgG and Cy3 goat anti-rabbit IgG (Invitrogen), Marina-Blue goat anti-mouse IgG (ThermoFisher), Alexa Fluor 488 goat anti-rabbit IgG, and Alexa Fluor 488 goat anti-mouse IgG (both Invitrogen).

Cell lines

HtTERT RPE-1 cells were purchased from ATCC (CRL-4000). Hek293 Flp-In TRex cells were purchased from Invitrogen (R78007). All other cell lines were gifts: Hek293T from M. W. Kirschner, HeLa-K from D. Gerlich, and securin knockout and parental HCT116 from C. Lengauer. Validation procedures for purchased cell lines are as described by the corresponding manufacturers. All other cell lines were authenticated via visual inspection of typical morphology, immunoblotting analyses (for example, absence of securin), cell synchronization behaviour,

efficiencies of different transfection reagents and resistance to certain antibiotics. Cell lines were not tested for mycoplasma contamination but microscopic inspections of their fluorescently labelled DNA contents were inconspicuous.

All human cells were cultured in Dulbecco's modified Eagle's medium (DMEM; Biowest) supplemented with 10% fetal calf serum (FCS; Sigma-Aldrich) at 37 °C and 5% CO₂. Generation of a stably transgenic Hek293-FlpIn-TRex line (Invitrogen) expressing Myc₆-separase(S1126A) upon induction with doxycycline has been described¹³. For transient expression of Flag₃-Tev₂-SGO2 variants (WT, R153D, N53I, M114R, RKK124–126A, LSE127–129A, HSDQ239–242A) Hek293T or HeLa-K cells were transfected with the corresponding pCS2-based plasmids using a calcium phosphate-based method or Lipofectamine 2000 (Invitrogen). For time-lapse experiments, HeLa cells stably expressing histone H2B-eGFP were used. In addition, H2B-mCherry-SCC1^{107–268}-eGFP¹⁹ was transiently transfected to visualize premature separase activation.

Cell treatments

For synchronization in early S-phase, cells were treated with 2 mM thymidine (Sigma-Aldrich) for 20 h. Synchronization of cells in prometaphase was done by addition of taxol (LC Laboratories) to 0.2 µg/ml 6 h after release from a single thymidine block. G2 arrest was achieved by addition of 10 µM RO-3306 (Santa-Cruz Biotechnology) 4 h after G1/S release. To analyse cells in G1-phase, cells were collected 15 h after release from a single thymidine block. For the 'taxol-ZM override' experiments, taxol-arrested mitotic HeLa-K cells were collected by shake-off and released for the indicated times by reseeding into medium supplemented with ZM-447439 (5 µM, Tocris Biosciences), taxol (0.2 µg/ml), cycloheximide (30 µg/ml, Sigma-Aldrich) and, where indicated, with proTame (6 µM, Boston Biochemicals) and Apcin (20 µM, Tocris Bioscience). To further enrich the endogenous separase-SGO2-MAD2 complex for later isolation, securin was depleted by RNAi.

Immunoprecipitation

We lysed 1 × 10⁷ cells with a dounce homogenizer in 1 ml LP2 lysis buffer (20 mM Tris-HCl pH 7.7, 100 mM NaCl, 10 mM NaF, 20 mM β-glycerophosphate, 5 mM MgCl₂, 0.1% Triton X-100, 5% glycerol), combined with benzonase (ad 30 U/l; Santa-Cruz Biotechnology), and incubated them on ice for 1 h. To preserve phosphorylation, lysis buffer was additionally supplemented with 50 nM calyculin A (LC-Laboratories) and 1 µM microcystin LR (Alexis Biochemicals). Corresponding lysates were centrifuged at 2,500g for 10 min followed by incubation of 1 ml cleared lysate with 10 µl of antibody-loaded beads for 4 h or overnight at 4 °C and washed 5× with LP2. In some cases, immobilized human SGO2 variants (purified from G1 cells) or endogenous separase-SGO2-MAD2 complex (purified from mitotic cells) were used as starting material for further experiments before bound proteins were eluted by boiling in SDS-sample buffer. For consecutive immunoprecipitation from *PTTG1*^{-/-} cells, the lysate used for the first purification was kept and served as origin for the second precipitation.

For immunoprecipitation of *X. laevis* separase, CSF-arrested *Xenopus* egg extract was prepared as previously described³⁴ and combined with cycloheximide (100 µg/ml), recombinant human Δ90-cyclin B1 (23 ng/µl ≈ 500 nM)¹⁹ and sperm nuclei (2,000 µl⁻¹). After 15 min at room temperature, the egg extract was released into anaphase II by addition of CaCl₂ (0.6 mM). Previously prepared mock- or *X. laevis* separase antibody-coupled magnetic beads were equilibrated in CSF-XB followed by addition of anaphase extract (minimal volume of 500 µl) and consecutive incubation for 45 min at 18 °C. After re-isolation, beads were washed 5× in CSF-XB supplemented with 300 mM NaCl and 0.01% Triton X-100.

RNA interference

For efficient knockdown, cells were calcium phosphate or RNAiMax (Invitrogen) transfected with 70–100 nM siRNA duplex

of *PTTG1*: 5'-UCUUAGUGCUUCAGAGUUUGUGUGUAU-3'; *SGO2*: 5'-GAA CACAUUUCUUCGCCUATT-3'; *MAD2*: 5'-GAGUCGGGACCACAGUUUA UU-3'; *MAD1*: 5'-AACCAGCGGCUCAAGGAGGUU-3'; *P31^{comet}*: 5'-GGCU GCUGUCAGUUUACUUTT-3'; *TRIP13*: 5'-CUGAUGAAGUGUCAGAUCA-3'; *TRP*: 5'-GGGUGAAGAUAGUAAUGAAUUCTT-3'. Transfected cells were grown for 12–24 h before synchronization procedures were applied. *Luciferase* siRNA (GL2) was used as negative control (Ctrl).

Immunofluorescence staining and microscopy

Hela-K cells transfected with the indicated siRNAs were grown on poly-lysine coated glass coverslips and processed 8 h after thymidine release in the presence of BI-2536 (10 nM) to slow down prophase in early mitotic cells. To remove soluble proteins, cells were pre-extracted (PBS, 0.3% Triton X-100) for 5 min, washed once with PBS and fixed with fixation solution (PBS, 3.7% formaldehyde, 0.3% Triton X-100) for 10 min at room temperature. Subsequently, coverslips were washed twice with quenching solution (PBS, 100 mM glycine), incubated with permeabilization solution (PBS, 0.5% Triton X-100) for 5 min, washed once with PBS and then incubated in blocking solution (PBS, 1% (w/v) BSA) for 1 h at room temperature. Coverslips were transferred into a wet chamber and incubated with primary antibodies for 1 h followed by four washes with PBS-Tx (PBS, 0.1% Triton X-100). After incubation with fluorescently labelled secondary antibodies for 40 min, samples were washed once, stained for 10 min with 1 µg/ml Hoechst 33342 in PBS-Tx and washed again four times. Finally, coverslips were mounted in 20 mM Tris-HCl pH 8.0, 2.33% (w/v) 1,4-diazabicyclo(2.2.2)octane, 78% glycerol on a glass slide. IFM of fixed cells was performed on a DMI 6000 inverted microscope (Leica) using a HCX PL APO 100×/1.40–0.70 oil objective. To identify early mitotic nuclei, DNA morphology (commencing condensation) and condensin staining intensity were examined. For representative images, Z-stack series over 4 µm in 0.2-µm increments were collected, deconvoluted and projected into one plane using the LAS-AF software. Chromosome spreads were prepared using Canoy's solution as described³⁵. Spreads were observed with the Zeiss Axioplan 2 Imaging microscope using a Plan-APOCHROMAT 100×/1.40 Oil objective. A cell was counted as suffering from PCS when exhibiting loss of cohesion of >50% of its chromosomes. At least 100 spreads were counted per condition. To assess centriole engagement status, centrosomes were isolated from RO-3306 arrested HeLa-K cells (4×10^6) 36 h after transfection of indicated siRNAs and stained as previously described³³.

Clonogenic assays

Twelve hours after siRNA transfection, HeLa-K, Hek293T, HCT116 parental or *PTTG1*^{-/-} cells were seeded in 10-cm dishes (100 cells per plate and condition). Cells were grown for 10 days and then fixed in ice-cold methanol for 10 min. Staining was performed as described³⁶ except for the use of ethanol instead of glutaraldehyde. The number of colonies per plate with a minimal area of 20 (circularity 0.00–1.00) was determined with ImageJ particle analysis software.

In vitro disassembly of separase–SGO2–MAD2 complex

Ten microlitres of immobilized separase–SGO2–MAD2 complex was incubated with 2 mg/ml recombinant variants of TRIP13 (WT; A, Walker A mutant G184A; B, Walker B mutant E253Q) and RGS-His₆–p31^{comet} (WT; QF, Q83A/F191A MAD2-binding deficient; PK, P228A/K229A TRIP13-binding deficient) in EDTA/EGTA-free lysis buffer supplemented with ATP (1 mM) in a final volume of 30 µl for 30 min at 18 °C. Subsequently, 15 µl of the corresponding supernatant was removed for later analysis. Beads were washed three times with LP2 and then equilibrated in cleavage buffer (10 mM Hepes-KOH pH 7.7, 50 mM NaCl, 25 mM NaF, 1 mM EGTA, 20% glycerol). To monitor separase activity, 2 µl of in vitro translated ³⁵S-RAD21–GFP was added to a volume of 30 µl. Following incubation for 30 min at 30 °C, reactions were stopped by boiling in SDS-sample buffer. Assaying disassembly by in vitro ubiquitylation

(Extended Data Fig. 10) was performed essentially as described³⁰ with the exception that instead of securin, 10 µl immobilized separase–SGO2–MAD2 complex was added as substrate.

Bacterially expressed proteins

pET28-vector encoded, His₆-SUMO₁-tagged³⁷ variants of TRIP13 (NP_004228) and p31^{comet} (NP_055443) were expressed individually in *E. coli* Rosetta 2 DE3 (Novagen). Bacteria were lysed in LP1 (PBS, 5 mM imidazole, 0.5 mM DTT and an additional 400 mM NaCl) and purified over Ni²⁺-NTA-agarose (Qiagen) according to standard procedures. Following elution with PBS supplemented with 250 mM imidazole, 0.5 mM DTT and an additional 400 mM NaCl (pH adjusted to 7.5 with HCl), proteins were dialysed at 4 °C against LP1 in presence of His₆-SENP2 (10 ng per 100 µg of protein) and then rotated for 3 h over 0.9× the amount of fresh Ni²⁺-NTA-agarose. Supernatants containing pure TRIP13 and p31^{comet} were dialysed against 50 mM Hepes-KOH pH 7.7, 10% glycerol, 100 mM NaCl, 5 mM MgCl₂, 1 mM EDTA and 1 mM DTT. Human MAD2 and *X. laevis* Mad2 were purified as described previously⁷.

H2A-peptide binding assay

Thr-9-phosphorylated and unmodified histone H2A-peptides (QAVLLP-KKTESHHKAKGK) were obtained from Bachem; Ser-10-phosphorylated or unmodified histone H3-peptides were purchased from Eurogentec (AS-61702 and AS-64611). One milligram of each of the H2A peptides was conjugated to 10 mg/ml maleimide-activated ovalbumin (ThermoFisher) according to the manufacturer's instructions and dialysed against CSF-XB³⁴ containing 0.5 mM DTT. Reactions (final volume of 30 µl) containing 10 µl immobilized FLAG₃-Tev₂-SGO2 and 1 µg ovalbumin-coupled-H2A peptide, 5 µg free H2A or H3 peptide, or 4 µg of recombinant MAD2 were assembled and incubated for 30 min at 18 °C. Unbound peptide or protein was removed by four washes with CSF-XB and the second reaction was assembled with the corresponding counterpart and again incubated for 30 min at 18 °C. Samples were again washed (four times with CSF-XB containing 0.01% Triton X-100) before beads were eluted by boiling in SDS-sample buffer.

Mapping experiments in *X. laevis* Sgo1

N-terminally tagged Myc₆- or FLAG₃-tagged *X. laevis* Sgo1 variants were in vitro translated in rabbit reticulocyte lysate (TNT Quick, Promega) according to the manufacturer's protocol. For mapping with the help of interaction-blocking antibodies, 16 µl Sgo1 was combined with 2.5 µg anti-*X. laevis* Sgo1 and 5 µg recombinant *X. laevis* Mad2 and incubated for 30 min at 18 °C. Then, 10 µl of magnetic beads loaded with *X. laevis* separase (isolated from anaphase egg extract) was added and reactions were incubated for 30 min at 18 °C. Beads were washed four times with CSF-XB, 0.01% Triton X-100 and twice with CSF-XB, 0.01% Triton X-100, 300 mM NaCl and finally eluted by boiling in SDS-sample buffer. In all other cases 12.5 µl Sgo1 and 5 µg Mad2 were pre-incubated in 300 µl CSF-XB, 0.01% Triton X-100 followed by addition of 10 µl immobilized separase.

Separase inhibition and cleavage assays

For separase inhibition, 10 µl of FLAG₃-Tev₂-SGO2 bound to anti-FLAG beads and 4 µg recombinant MAD2 were incubated in cleavage buffer (total volume of 28 µl) for 10 min at 18 °C followed by the addition of 2.5 µl active (Ac, S1126A variant) or protease dead (PD, S1126A and C2029S) separase¹⁷. After a 10-min incubation at room temperature, 2 µl in vitro translated, ³⁵S-labelled RAD21–GFP were added. Reactions were stopped after 30 min by boiling in SDS-sample buffer. Samples were separated by SDS–PAGE and blotted onto PVDF membrane (SERVA), which was cut and analysed by immunoblotting before reassembly and autoradiography. For Fig. 3b, 2 µl of each in vitro translated, ³⁵S-labelled SGO2 variant was combined with 1 µg human MAD2 and, where indicated, 0.1 µg PLK1 (ProKinase; No. 0183-0000-1) in modified cleavage buffer (10 mM Hepes-KOH pH 7.7, 50 mM NaCl, 25

Article

mM NaF, 20% glycerol, 1 mM ATP, 10 mM MgCl₂) and a total volume of 15 µl. After 15 min at 30 °C, 2.5 µl separase (Ac or PD) was added and reactions were incubated for 30 min at room temperature. For Extended Data Fig. 9a, 2 µl of each in vitro translated, ³⁵S-labelled *X. laevis* Sgo1 variant was combined with 1 µg *X. laevis* Mad2 (WT or ΔC10) and 10 µl immobilized *X. laevis* separase. Samples were further processed and analysed as described above.

Live-cell imaging

Cells in phenol red-free medium were seeded into µ-slide 8-well chambered coverslips (Ibidi) and kept in an atmosphere of 37 °C and 5% humidified CO₂ during microscopy on a DMI 6000 inverted microscope (Leica). For imaging of unperturbed mitosis, GFP and DIC images were captured 6 h after release from thymidine arrest at 10-min intervals over a period of 15 h, through a HCX PL APO 40×/0.85 CORR objective. Changes in focus plane due to mitotic rounding of the cells were compensated by collecting Z-stacks at each time point. Captured images from each experiment were analysed using the corresponding LAS-AF software (Leica).

Statistics and reproducibility

No statistical methods were used to predetermine sample size. The experiments were not randomized. For quantitative analyses of chromosome spreads, clonogenic assays, and IFM specimen the investigators were blinded to sample allocation. Experiments analysed by immunoblotting were repeated 2–4 times with similar results (2–4 biological replicates).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All source data for this study are available online.

30. Hellmuth, S., Böttger, F., Pan, C., Mann, M. & Stemmann, O. PP2A delays APC/C-dependent degradation of separase-associated but not free securin. *EMBO J.* **33**, 1134–1147 (2014).
31. Gorr, I. H. et al. Essential CDK1-inhibitory role for separase during meiosis I in vertebrate oocytes. *Nat. Cell Biol.* **8**, 1035–1037 (2006).
32. Hellmuth, S., Gutiérrez-Caballero, C., Llano, E., Pendás, A. M. & Stemmann, O. Local activation of mammalian separase in interphase promotes double-strand break repair and prevents oncogenic transformation. *EMBO J.* **37**, e99184 (2018).
33. Schöckel, L., Möckel, M., Mayer, B., Boos, D. & Stemmann, O. Cleavage of cohesin rings coordinates the separation of centrioles and chromatids. *Nat. Cell Biol.* **13**, 966–972 (2011).
34. Murray, A. W. Cell cycle extracts. *Methods Cell Biol.* **36**, 581–605 (1991).
35. McGuinness, B. E., Hirota, T., Kudo, N. R., Peters, J. M. & Nasmyth, K. Shugoshin prevents dissociation of cohesin from centromeres during mitosis in vertebrate cells. *PLoS Biol.* **3**, e86 (2005).
36. Franken, N. A., Rodermond, H. M., Stap, J., Haveman, J. & van Bree, C. Clonogenic assay of cells in vitro. *Nat. Protocols* **1**, 2315–2319 (2006).
37. Butt, T. R., Edavettal, S. C., Hall, J. P. & Mattern, M. R. SUMO fusion technology for difficult-to-express proteins. *Protein Expr. Purif.* **43**, 1–9 (2005).
38. Holland, A. J., Böttger, F., Stemmann, O. & Taylor, S. S. Protein phosphatase 2A and separase form a complex regulated by separase autocleavage. *J. Biol. Chem.* **282**, 24623–24632 (2007).
39. Schägger, H. Tricine-SDS-PAGE. *Nat. Protocols* **1**, 16–22 (2006).

Acknowledgements We thank S. Heidmann and P. Wolf for critical reading of the manuscript, and J. Hübner and M. Hermann for technical assistance. This work was supported by a grant (STE997/4-2) from the Deutsche Forschungsgemeinschaft (DFG) to O.S. and by MINECO (BFU2017-89408-R) and Junta de Castilla y León (CSI239P18). CIC-IBMCC is supported by the Programa de Apoyo a Planes Estratégicos de Investigación de Estructuras de Investigación de Excelencia cofunded by the Castilla–León autonomous government and the European Regional Development Fund (CLC–2017–01).

Author contributions L.G.-H. and A.M.P. first discovered the interaction between separase and SGO2. S.H. carried out all experiments except for the one shown in Extended Data Fig. 1a, which was conducted by L.G.-H. S.H. and O.S. co-designed the research and wrote the paper.

Competing interests The authors declare no competing interests.

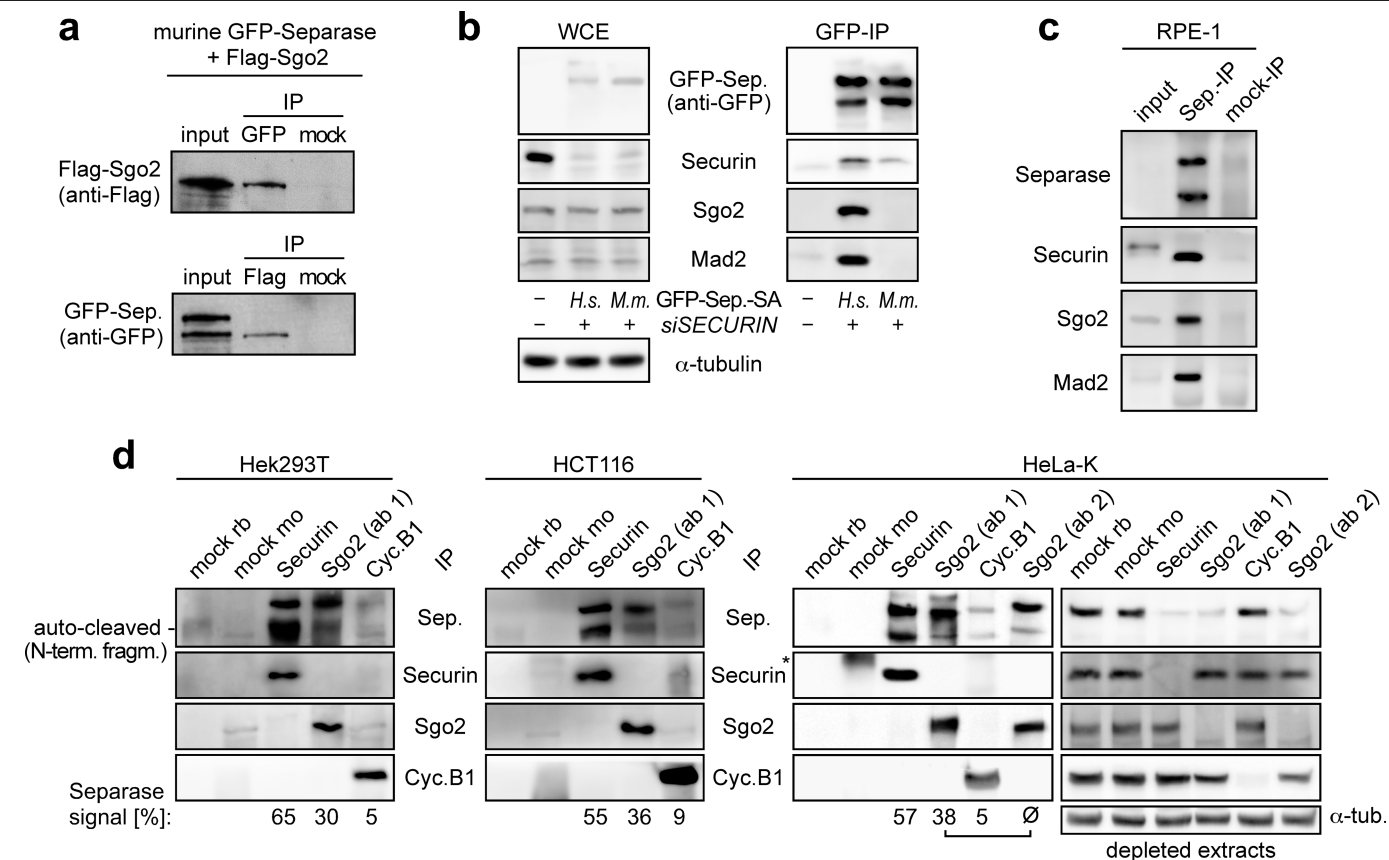
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2182-3>.

Correspondence and requests for materials should be addressed to O.S.

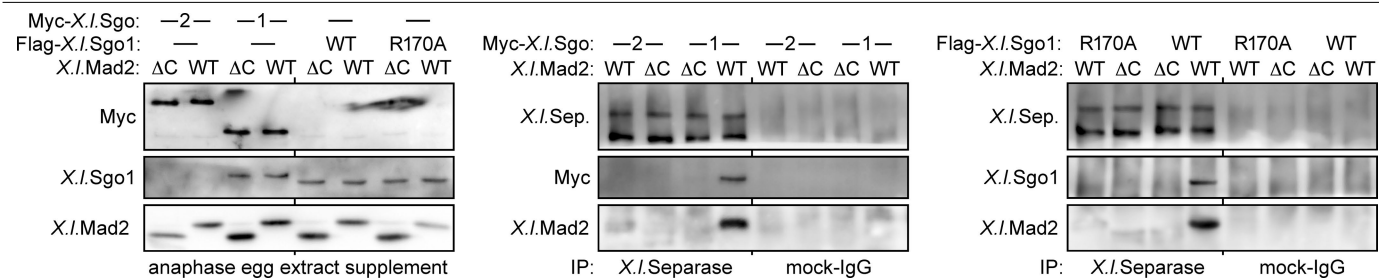
Peer review information Nature thanks Silke Hauf, Adele Marston and Hongtao Yu for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



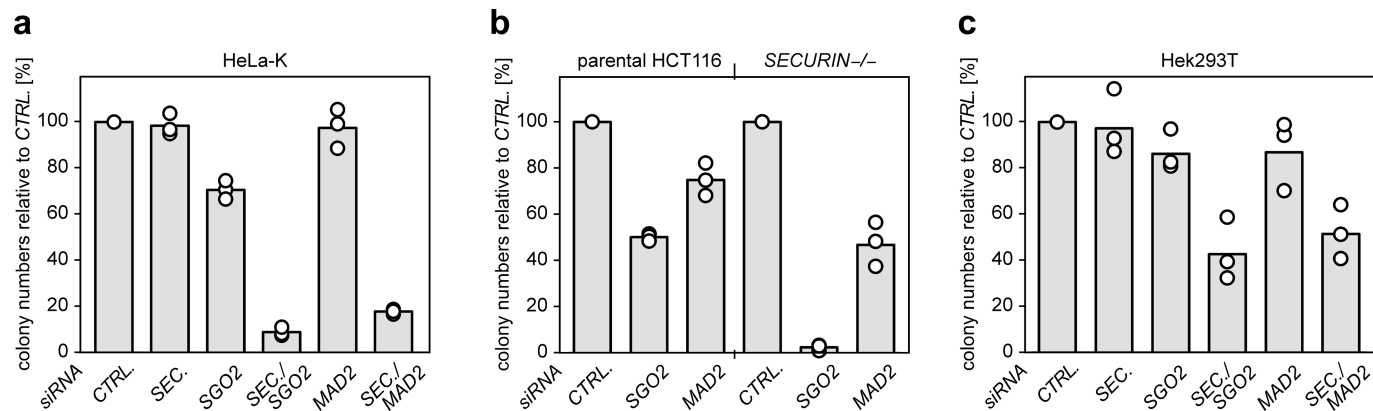
Extended Data Fig. 1 | Mammalian SGO2 interacts with separase. **a**, Hek293T cells were co-transfected with expression vectors for the following mouse proteins: GFP-separase, Flag-SGO2 and securin. Immunoprecipitation was carried out from transfected cells with either anti-Flag or anti-GFP antibodies and analysed by immunoblotting with the indicated antibodies. **b**, Mouse separase does not interact with human SGO2 and, therefore, cannot be used to study separase regulation in human cells. Securin-depleted Hek293T cells expressing GFP-tagged human separase(S1126A) or mouse separase(S1121A) were taxol-arrested and then subjected to IP-immunoblotting analyses using the indicated antibodies. **c**, SGO2 and MAD2 interact specifically with separase in untransformed cells. Taxol-arrested RPE1 cells were subjected to IP-

immunoblotting analyses as indicated. **d**, Even in securin-expressing cells, a considerable fraction of separase is sequestered by SGO2. Taxol-arrested Hek293T, HCT116, and HeLa-K cells were subjected to IP-immunoblotting analyses using the indicated antibodies. ab 1, anti-DVPPRESHSHSDQSSKC (corresponding to amino acids 230-245 of human SGO2); ab 2, anti-KSEDLSERTSRRRRC (corresponding to amino acids 1,234-1,249 of human SGO2); rb, rabbit; mo, mouse. Given below are the relative intensities (in per cent) of the separase signals (sum of full-length and N-terminal auto-cleavage fragment). Note the considerable co-depletion of separase upon SGO2 immunoprecipitation from HeLa-K (right).

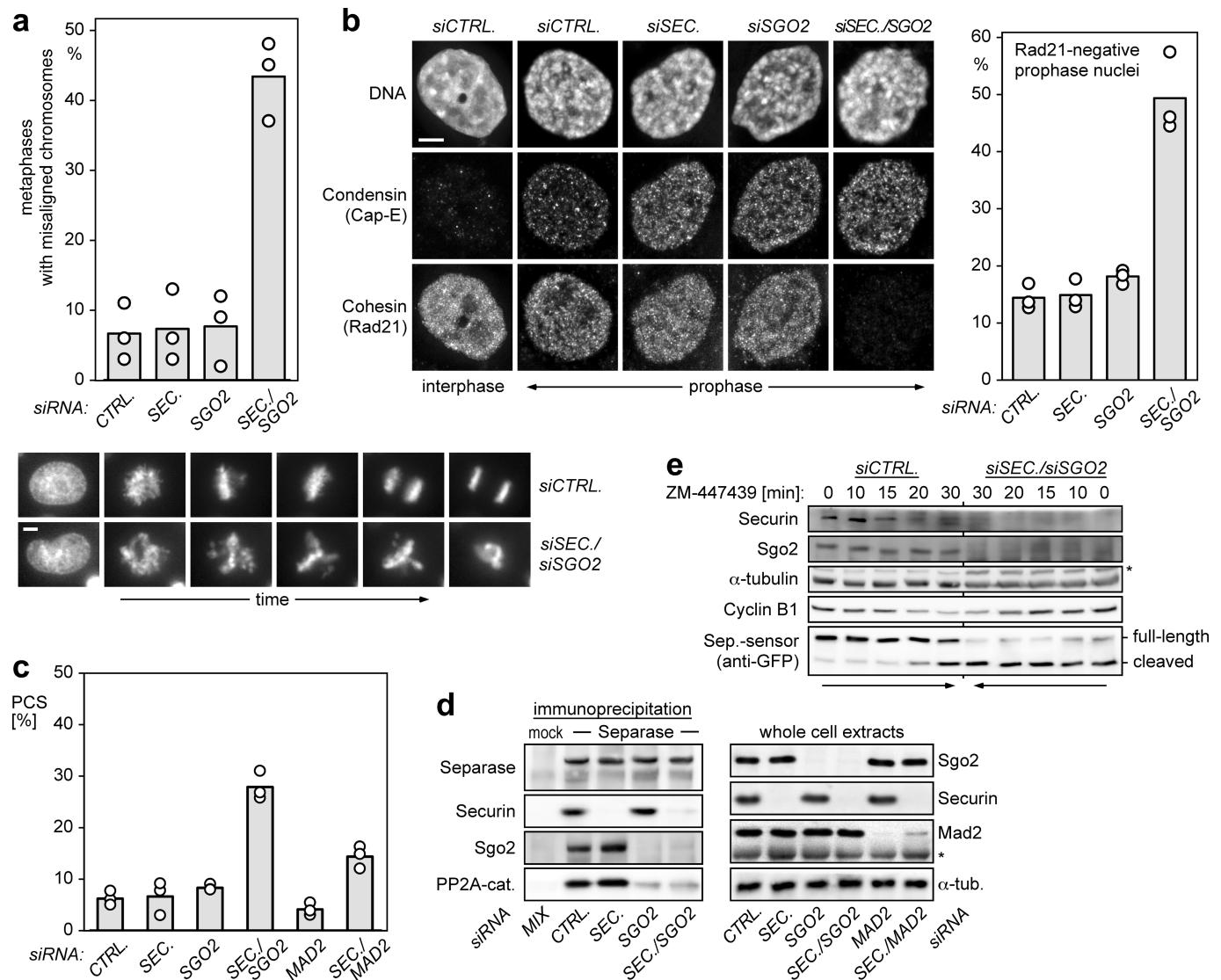


Extended Data Fig. 2 | Sgo1 rather than Sgo2 interacts with Mad2 and separase in *Xenopus*. The indicated variants of in vitro-expressed *X. laevis* (*X.l.*) shugoshins and an excess of *E. coli*-expressed Mad2 (to mimic SAC signalling) were incubated in anaphase egg extract (left). Following IP with

anti-*X.l.* separase or mock-IgG from these mixtures, isolated proteins were detected by immunoblotting (middle and right). R170A, Mad2-binding-deficient Sgo1; ΔC, Sgo1-binding-deficient, C-terminally truncated Mad2.

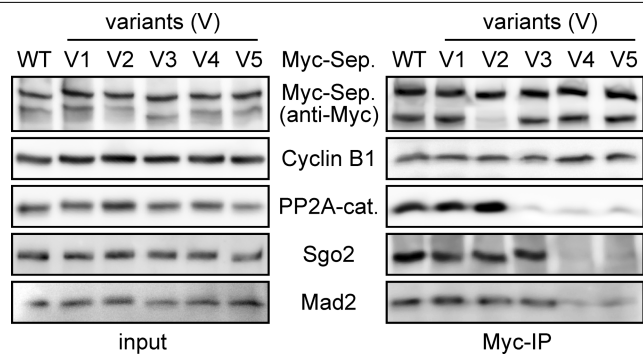


Extended Data Fig. 3 | Lack of securin and SGO2 or MAD2 have synergistic cytotoxic effects. a–c, Clonogenic assays with HeLa-K (**a**), *PTTG1*^{-/-} (*SECURIN*^{-/-}) and parental HCT116 cells (**b**), and Hek293T cells (**c**) transfected with the indicated siRNAs. Bars show percentages of colony numbers relative to the control of three independent experiments (dots).



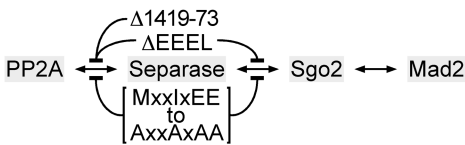
Extended Data Fig. 4 | Depletion of securin and SGO2, but not the individual knockdowns, results in impaired chromosome alignment, premature loss of chromosomal cohesin and unscheduled separase activity. **a**, HeLa-K cells transfected with the indicated siRNAs and a histone H2B-eGFP expression plasmid were observed by video fluorescence microscopy to assess metaphase plate formation (bottom). Bars show mean of three independent experiments (dots) counting at least 50 mitotic cells each (top). Scale bar, 5 μ m. **b**, Eight hours after release from thymidine arrest, HeLa-K cells transfected with the indicated siRNAs were pre-extracted, fixed, and examined by fluorescence microscopy for cohesin-negative early mitotic chromatin. Left, representative images; right, bars show mean of three

independent experiments (dots) counting prophase nuclei that were still round but already stained positive for condensin (100 each). Scale bar, 5 μ m. **c**, Hek293T cells transfected with the indicated siRNAs were supplemented with taxol (but not EGCG; compare Fig. 2b) and analysed by chromosome spreading. Bars show mean of three independent experiments (dots). **d**, Exemplary immunoblots of cells analysed in **c** and Fig. 2b. Star denotes nonspecific band. **e**, siRNA-transfected HeLa-K cells expressing a separase activity sensor (H2B-mCherry-RAD21¹⁰⁷⁻²⁶⁸-eGFP) were released from a taxol arrest by addition of ZM-447439 at time zero and analysed by time-resolved immunoblotting.

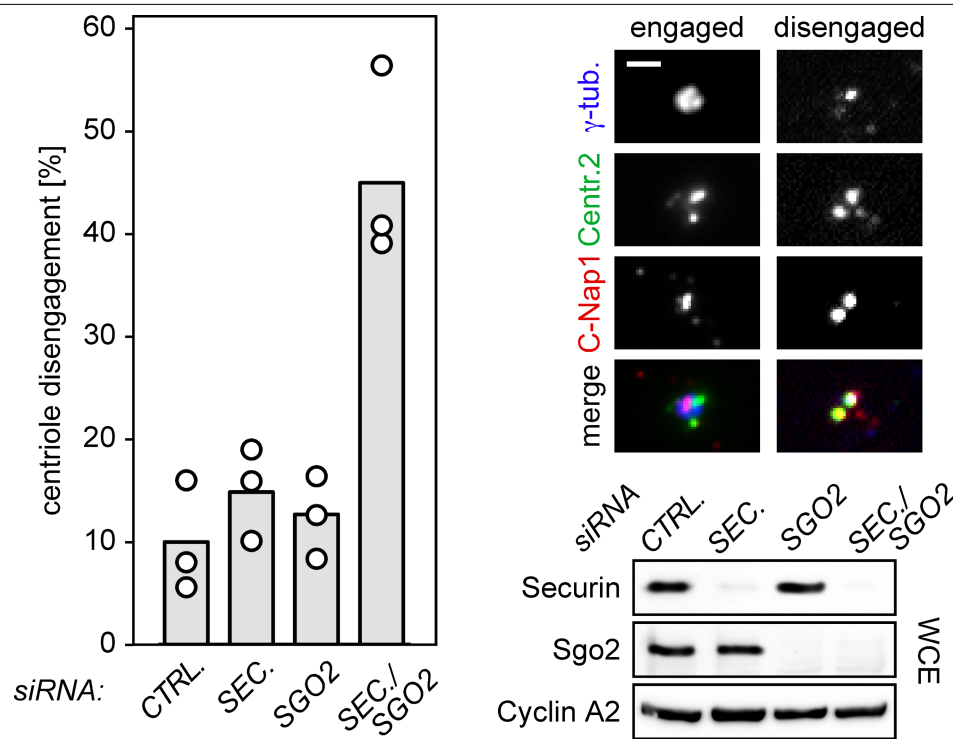


Extended Data Fig. 5 | Identification of SGO2-binding-deficient separase variants. Taxol-arrested Hek293T cells expressing transgenic Myc-tagged wild-type separase (WT) or one of the indicated variants (V1-V5) were analysed by IP-immunoblotting using the indicated antibodies. The investigation of

variants:	changes:	amino acid positions:
V1 (ABBA)	FxVFXE to AxAAxA	1362-7
V2 (non-cleavable):	ExxR to RxxE	1483-6, 1503-6, 1532-5
V3 (PP2A-1):	$\Delta 55$	1419-73
V4 (PP2A-2):	$\Delta EEEEL$	1490-3
V5 (PP2A-3):	MxxIxEE to AxxAxAA	1485-91

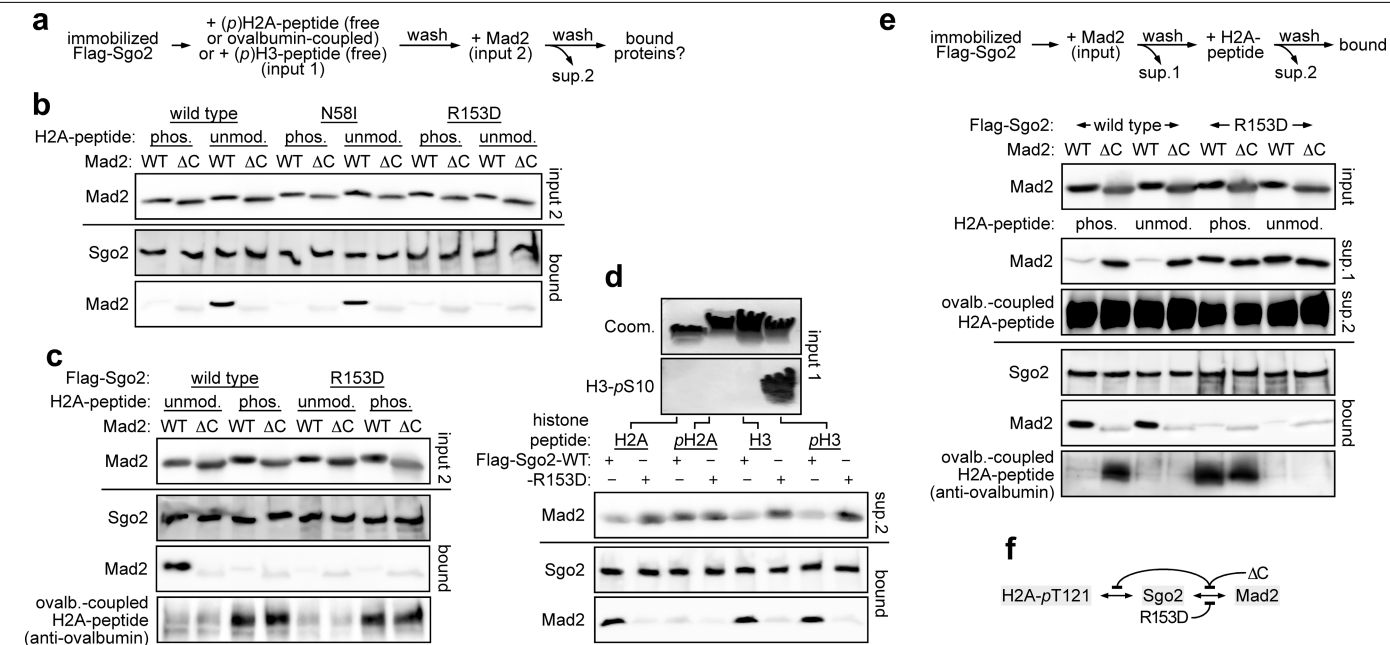


PP2A-binding-deficient variants was motivated by the notion that SGO2-MAD2, like PP2A, preferentially interacts with full-length rather than auto-cleaved separase³⁸ (see, for example, Fig. 3c). V3 cannot bind PP2A but can bind SGO2-MAD2, indicating overlapping but not identical binding sites.



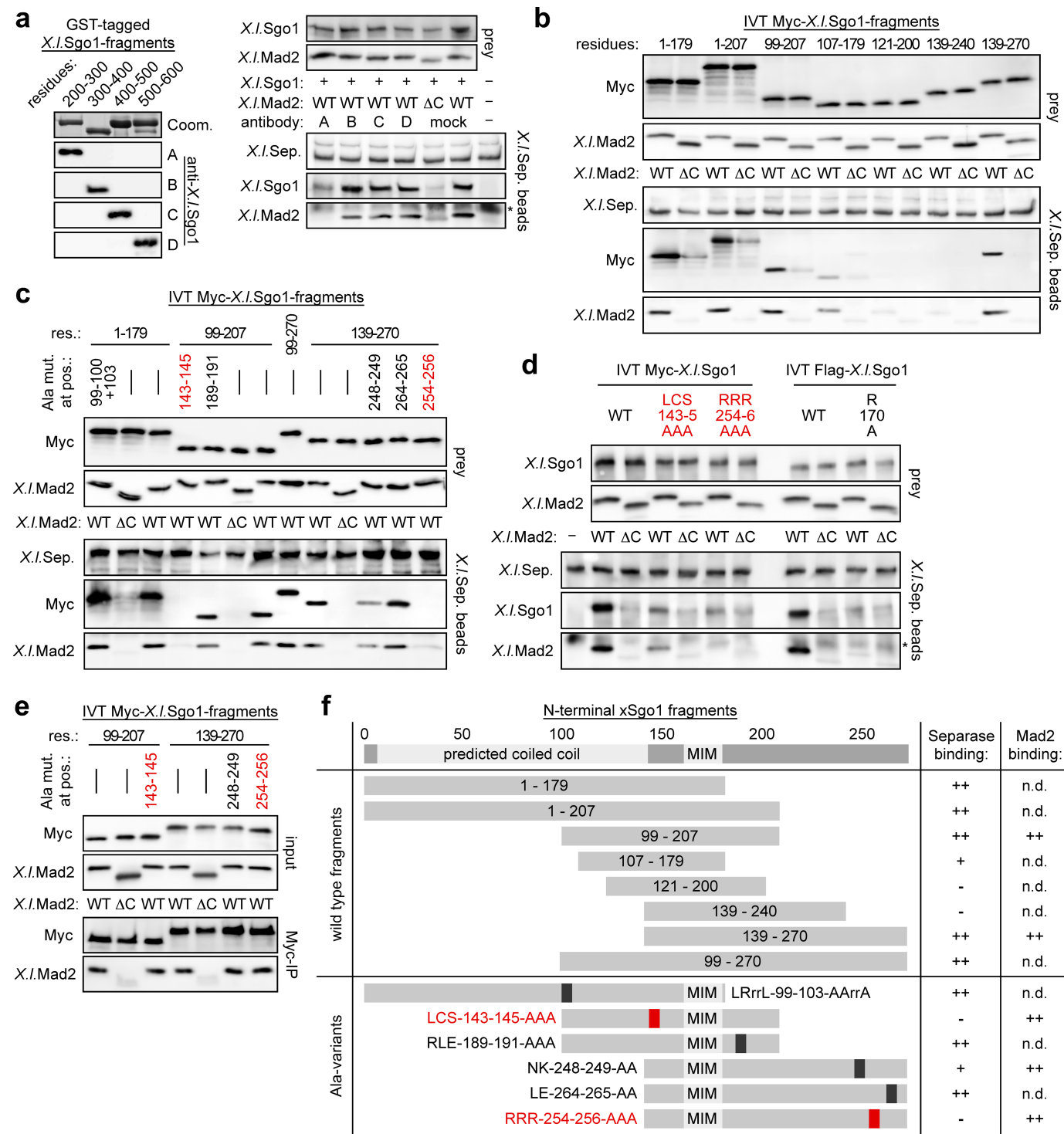
Extended Data Fig. 6 | Depletion of securin and Sgo2 but not the individual knock-downs result in premature disengagement of centrioles. HeLa-K cells transfected with the indicated siRNAs were released from a thymidine block and arrested in G2 phase with the CDK1 inhibitor RO-3306. Corresponding lysates were used for immunoblotting (bottom right) and

centrosome isolation followed by immunofluorescence microscopy (top right, representative images) to assess the degree of centriole disengagement as revealed by two C-Nap1 foci. Left, bars show mean of three independent experiments (dots) counting 100 centrosomes each. Scale bar, 1 μ m.



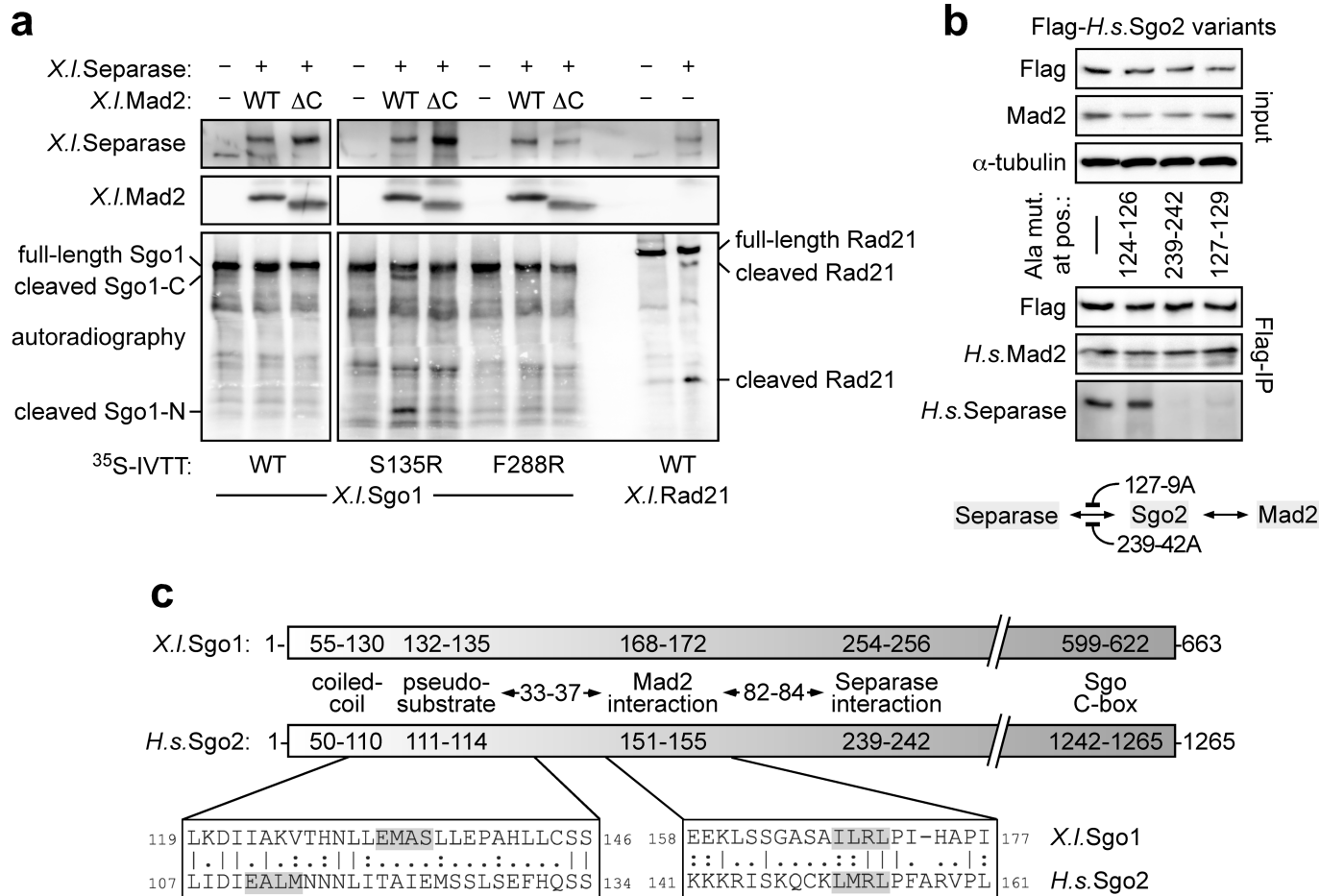
Extended Data Fig. 7 | Mutually exclusive binding of SGO2 to MAD2 or phosphorylated histone H2A. **a–d**, Pre-charging of SGO2 with phosphorylated histone H2A blocks subsequent MAD2 binding. **a**, Experimental scheme for experiments shown in **b–d**. Beads loaded with the indicated Flag-SGO2 variants were consecutively incubated first with phosphorylated or unphosphorylated H2A or H3 peptide (input 1) and then with wild-type or, where indicated, C-terminally truncated (ΔC) MAD2 (input 2). Following washing (supernatant 2), bound proteins were visualized by immunoblotting. **b**, Usage of free H2A peptides. **c**, Usage of ovalbumin-coupled

H2A peptides to facilitate their detection by standard glycine-SDS-PAGE and immunoblotting. **d**, Phosphorylated histone H3 peptide does not bind to SGO2. Free H2A and H3 peptides, used to interrogate immobilized SGO2, were separated by Tricine-SDS-PAGE³⁹ and analysed by Coomassie staining and immunoblotting. **e**, Pre-charging of SGO2 with MAD2 blocks subsequent phospho-H2A binding. Experimental scheme (top) and corresponding immunoblotting analysis (bottom). H2A peptides were coupled to ovalbumin for ease of detection. **f**, Cartoon illustrating that SGO2 can bind to H2A phosphorylated on Thr121 or to MAD2 but not both at the same time.



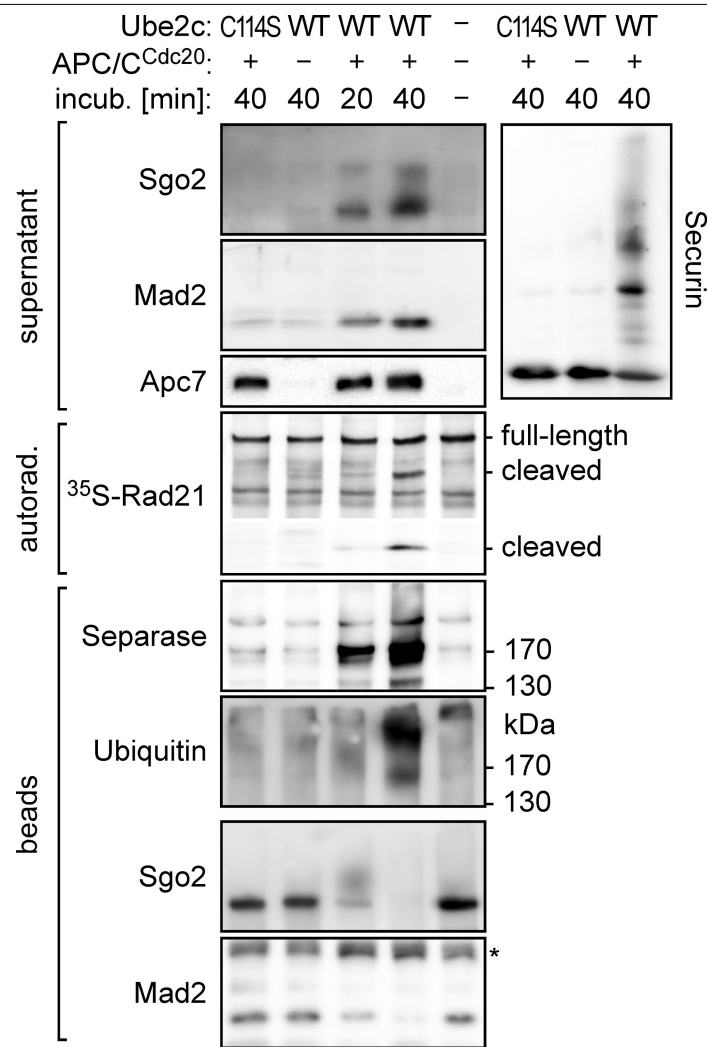
Extended Data Fig. 8 | Mapping separase-binding sites on *X. laevis* Sgo1. a, Covering amino acids 200–300 of *X. laevis* Sgo1 with polyclonal antibodies impairs its binding to separase. Region-specific polyclonal *X. laevis* Sgo1 antibodies A–D were characterized (left) and added to in vitro-expressed *X. laevis* Sgo1 and *E. coli*-expressed MAD2 (prey) as indicated. Immobilized *X. laevis* separase isolated by immunoprecipitation from anaphase egg extracts was then added as bait to these mixtures. Separase beads were washed and finally probed for associated proteins by immunoblotting. Mock, unspecific IgG; asterisk, unspecific band. **b, c,** Identification of two sites within *X. laevis* Sgo1 that are relevant for separase binding. Different in vitro translated (IVT) *X. laevis* Sgo1 fragments and variants thereof were combined with wild-type

MAD2 or MAD2-ΔC as indicated (prey). Separase beads as in **a** were combined with these mixtures, washed and analysed for associated proteins by immunoblotting. **d–f,** *X. laevis* Sgo1(143–145A) and Sgo1(254–256A) (red) show compromised separase binding but retain MAD2 binding. **d,** The indicated full-length *X. laevis* Sgo1 variants were assessed for MAD2-dependent separase binding as in **b, c, e.** The indicated *X. laevis* Sgo1 fragments and variants thereof were combined with wild-type MAD2 or MAD2-ΔC, immunoprecipitated via their Myc-tag and assessed for MAD2 binding by immunoblotting. **f,** Summary of the mapping experiments. MIM, Mad2-interaction motif; n.d., not determined.



Extended Data Fig. 9 | *X. laevis* Sgo1 and human SGO2 share the same order and spacing of separase- and MAD2-binding sites. **a**, A point mutation turns *X. laevis* Sgo1 into a Mad2-dependent separase substrate. ³⁵S-labelled *X. laevis* Sgo1 variants were incubated with wild-type Mad2 or Mad2-ΔC before being assayed for in vitro-cleavage by *X. laevis* separase, which had been immunoprecipitated from anaphase egg extracts. Gels were blotted onto membranes, which were cut and subjected to immunoblotting (top) before reassembly and autoradiography (bottom). **b**, Human (*H.s.*) SGO2(127-129A)

and SGO2(239-242A) show compromised separase binding but retain MAD2 binding. The indicated full-length SGO2 variants were immunoprecipitated via their Flag-tags from transfected, taxol-arrested Hek293T cells and assessed by immunoblotting for binding of MAD2 and separase. See also illustration at bottom. **c**, Cartoon comparing the arrangement of functional domains and alignment of pseudo-substrate and Mad2-binding sites of *X. laevis* Sgo1 and human SGO2. The separase interaction site around position 144 of *X. laevis* Sgo1 and around position 128 of human SGO2 was omitted for clarity.



Extended Data Fig. 10 | In vitro disassembly of separase–SGO2–MAD2 by APC/C^{CDC20}-dependent ubiquitylation. Immobilized separase–SGO2–MAD2 complex isolated as in Fig. 4a was incubated with ATP, ubiquitin, E1, UBE2S, wild-type or dominant-negative (DN) UBE2C and APC/C^{CDC20} (or reference buffer). After removal of the supernatant and washing, the beads were

incubated with ³⁵S-RAD21 before being analysed by autoradiography and immunoblotting. The autoradiograph shows the relevant upper and lower parts of the same gel. Ubiquitylation of purified securin served as a positive control.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☐ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☐ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection No commercial open source software was used to collect data.

Data analysis Because no complex statistical analysis is performed standard analysis programmes (excel 2016, sigma Plot 9.0 or ImageJ 1.41) were used.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The authors declare that all data supporting the findings of this study are available within the paper and its supplementary information files. If there is reasonable request data can also be provided from the corresponding author.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- ☒ Life sciences
- ☐ Behavioural & social sciences
- ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We did not perform statistical analysis to determine a specific sample size. We used standard cell culture based experiments and our sample sizes represent those generally used in this field of research. Clonogenic assays and experiments involving chromosome spreads or IFM were independently repeated at least three times. Experiments analysed by immunoblotting were repeated 2-4 times with similar results. Within one sub-figure (a, b, c,...), all shown immunoblots stem from one and the same, representative experiment.
Data exclusions	No data was excluded.
Replication	The values obtained from distinct experimental trials were reproducible. The data are presented as means together with the corresponding individual data points from each repetition to indicate biological variation.
Randomization	For quantitative analyses of chromosome spreads, clonogenic assays, and IFM specimen the investigators were blinded to sample allocation.
Blinding	For quantitative microscopic analysis of chromosome spreads the sample identity was blinded.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems	Methods
<div><div>n/a</div><div>Involved in the study</div><div><div><input type="checkbox"/></div><div><input checked="" type="checkbox"/></div>Antibodies</div><div><div><input type="checkbox"/></div><div><input checked="" type="checkbox"/></div>Eukaryotic cell lines</div><div><div><input checked="" type="checkbox"/></div><div><input type="checkbox"/></div>Palaeontology</div><div><div><input checked="" type="checkbox"/></div><div><input type="checkbox"/></div>Animals and other organisms</div><div><div><input checked="" type="checkbox"/></div><div><input type="checkbox"/></div>Human research participants</div><div><div><input checked="" type="checkbox"/></div><div><input type="checkbox"/></div>Clinical data</div></div>	

n/a

Involved in the study

☒

☐

ChIP-seq

☒

☐

Flow cytometry

☒

☐

MRI-based neuroimaging

Antibodies

Antibodies used	Rabbit anti-separase (1:1,500) see Ref. in Material and Methods, mouse anti-securin (1:1,000; clone DCS-280; Code No. K009-3; MBL), mouse anti-Flag M2 (1:2,000; Product No. F1804; Sigma-Aldrich), rabbit anti-Sgo2 (1:1,000; Product No. A301-262A; Lot. No. A301-262A-1; Bethyl), rabbit or guinea pig anti-Sgo2 (1µg/ml; raised by Charles river laboratories against the peptide: DVPPRESHSQSSKC), rabbit anti-Sgo1 (1:500, Abcam ab21633), mouse anti-Mad2 (1:800; clone 17D10; Product No. sc-47747; Santa Cruz Biotechnology), rabbit anti-Mad2 (1:1,000; Product No. A300-300A; Bethyl), mouse anti-Mad1 (1:1,000; clone 9B10; Product No. M8069; Sigma); guinea pig anti-TRIP13 (1.5 µg/ml; raised by Charles River Laboratories against full length Trip13), rabbit anti-p31comet (2µg/ml; raised by Charles River Laboratories against isoform 2 of full length p31comet), mouse anti-APC7 (1:800; Product No. PA5-20948; Thermo Scientific), rabbit anti-phosphoSer10-histone H3 (1:1,000; Product No. 06-570; Lot No. 2370127; Millipore), mouse anti-PP2A-C (1:1,000; clone 1D6; Product No. 05-421; Millipore), mouse anti-cyclin B1 (1:1,000; Product No. 05-373; Lot. No. 2199734; Millipore), goat anti-Cdc27 (1:1000) see Ref. in Material and Methods, mouse anti-topoisomerase IIα (1:1,000; clone 1C5; Product No. ADI-KAM-CC210-E; Enzo Life Sciences), mouse anti-cyclin A2 (1:200; clone 46B11; Product No. sc-53234; Santa Cruz Biotechnology), mouse anti-Rad21 (1:800; clone B-2; Product No. sc-271601; Santa Cruz Biotechnology), rabbit anti-Rad21 (1:1,000; Product No. A300-080A; Bethyl). Antibodies directed against Xenopus proteins: Four different rabbit anti-Sgo1 (all 1 µg/ml used in WB; raised against amino acids 200-300, 300-400, 400-500, and 500-600 of X.l. Sgo1) see Ref. in Material and Methods, rabbit anti-separase (1:1000) see Ref. in Material and Methods, rabbit anti-Mad2. Other antibodies: Mouse anti-Myc (hybridoma supernatant 1:50; DSHB, 9E10), rat anti-HA (1:2,000; clone 3F10; Product No. 11867423001; Roche), rabbit anti-ovalbumin (1:1,000; Product No. PA1-196; Thermo Fisher Scientific), mouse anti-ubiquitinated proteins (1:1,000; clone FK2; Product No. 04-263; Lot No. 3117294; Millipore), mouse anti-GFP (hybridoma supernatant 1:2,000; gift from D. van Essen and S. Sacconi), and mouse anti-α-tubulin (hybridoma supernatant 1:200; DSHB; clone 12G10).
Validation	Antibodies are validated using multiple methods (e. g. ICC/IF, WB and IP). To address antibody specificity we used the corresponding recombinantly expressed antigen as positive- and RNAi-mediated depletion from human cells as negative controls.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)	hTERT RPE-1 cells were purchased from ATCC (CRL-4000). Hek293 Flp-In TRex cells were purchased from Invitrogen (R78007). All other cell lines were gifts: Hek293T from Marc W. Kirschner (Harvard Medical School), HeLa-K from Daniel Gerlich (IMBA), SECURIN knock-out and parental HCT116 from Christoph Lengauer (Johns Hopkins).
Authentication	Validation procedures for purchased cell lines are described by the corresponding manufacturers. All other cell lines were authenticated via visual inspection of typical morphology, by immunoblotting analyses (e.g. absence of securin), cell synchronization behavior, efficiencies of different transfection reagents and resistance to certain antibiotics.
Mycoplasma contamination	Cell lines were not tested for mycoplasma contamination but microscopic inspections of their fluorescently labeled DNA contents were inconspicuous.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Separase-triggered apoptosis enforces minimal length of mitosis

<https://doi.org/10.1038/s41586-020-2187-y>

Susanne Hellmuth¹ & Olaf Stemmann¹✉

Received: 22 March 2019

Accepted: 10 February 2020

Published online: 8 April 2020

 Check for updates

Prolonged mitosis often results in apoptosis¹. Shortened mitosis causes tumorigenic aneuploidy, but it is unclear whether it also activates the apoptotic machinery². Separase, a cysteine protease and trigger of all eukaryotic anaphases, has a caspase-like catalytic domain but has not previously been associated with cell death^{3,4}. Here we show that human cells that enter mitosis with already active separase rapidly undergo death in mitosis owing to direct cleavage of anti-apoptotic MCL1 and BCL-XL by separase. Cleavage not only prevents MCL1 and BCL-XL from sequestering pro-apoptotic BAK, but also converts them into active promoters of death in mitosis. Our data strongly suggest that the deadliest cleavage fragment, the C-terminal half of MCL1, forms BAK/BAX-like pores in the mitochondrial outer membrane. MCL1 and BCL-XL are turned into separase substrates only upon phosphorylation by NEK2A. Early mitotic degradation of this kinase is therefore crucial for preventing apoptosis upon scheduled activation of separase in metaphase. Speeding up mitosis by abrogation of the spindle assembly checkpoint results in a temporal overlap of the enzymatic activities of NEK2A and separase and consequently in cell death. We propose that NEK2A and separase jointly check on spindle assembly checkpoint integrity and eliminate cells that are prone to chromosome missegregation owing to accelerated progression through early mitosis.

The intrinsic pathway of apoptosis is regulated by a balance between pro- and anti-apoptotic BCL2 family proteins that are hallmarked by presence of one to four BCL2 homology (BH) domains⁵. Pore formation by homo-oligomerization of BAK and BAX leads to mitochondrial outer membrane permeabilization (MOMP) and release of cytochrome *c* and other apoptogenic factors from the intermembrane space. MOMP is counteracted by family members such as BCL2 itself, BCL2-like 1 (BCL-XL) and myeloid cell leukaemia 1 (MCL1). These proteins use a hydrophobic groove formed by their BH1–3 domains to sequester the BH3 domain of BAK/BAX and inhibit their self-interaction. BH3-only proteins such as BIM or BAD activate BAK/BAX either directly by transient interaction or indirectly by forcing BAK/BAX off anti-apoptotic BCL2 members through competition⁶. Intrinsic apoptosis in response to excessive cellular stress, such as DNA damage, is initiated by activation of BH3-only proteins, typically via upregulated transcription^{5,7}. Other triggers of intrinsic apoptosis are less well understood. Likewise, it remains unclear whether proteins other than BAK/BAX might also be able to form pores and contribute to MOMP.

Separase is the essential trigger protease of all eukaryotic anaphases⁴. Once activated in metaphase, it opens the DNA-embracing cohesin ring complex by cleavage of the kleisin subunit, thus resolving sister chromatid cohesion and enabling chromosome segregation. Separase contains a C-terminal caspase-like proteolytic domain³, but it has not been functionally linked to apoptosis. For most of the cell cycle, spindle assembly checkpoint (SAC) signalling ensures that human separase is held inactive by association with securin, SGO2–MAD2 or CDK1–cyclin B1^{8,9}. In response to improper attachment of kinetochores to spindle

microtubules, the SAC delays activation of separase and other late mitotic events, thereby giving the cell time for error correction^{2,10}. Hyperstimulation of the SAC by spindle toxins such as taxol (paclitaxel) prolongs mitosis¹¹. Conversely, SAC impairment results in shortened mitosis, chromosomal instability and tumorigenesis². While prolonged mitosis is known to result in death in mitosis (DiM) or mitotic slippage followed by apoptosis in interphase¹, it remains unstudied whether shortened mitosis might also trigger intrinsic apoptosis.

DiM upon premature separase activation

Both depletion of the cohesion-protecting factors SGO1 or sororin and derepression of separase (by co-depletion of SGO2 and securin) result in premature sister chromatid separation^{9,12,13} (Extended Data Fig. 1a). This was followed by prolonged mitotic arrest of cells lacking SGO1 or sororin, as previously described^{12,13}. However, cells lacking SGO2 and securin exhibited DiM as judged by cleavage of poly(ADP-ribose) polymerase (PARP) and fluorogenic caspase reporters (Extended Data Fig. 1b–h). DiM in cells depleted of SGO2 and securin was specific because it was suppressed by concomitant knockdown of separase (Extended Data Fig. 1g, h). Thus, premature activity of separase rather than premature sister chromatid separation represents an apoptotic stimulus.

MCL1 and BCL-XL are separase substrates

We speculated that MCL1 and BCL-XL could be relevant targets of separase in DiM because both (1) have previously been linked to apoptosis

¹Chair of Genetics, University of Bayreuth, Bayreuth, Germany. ✉e-mail: olaf.stemmann@uni-bayreuth.de

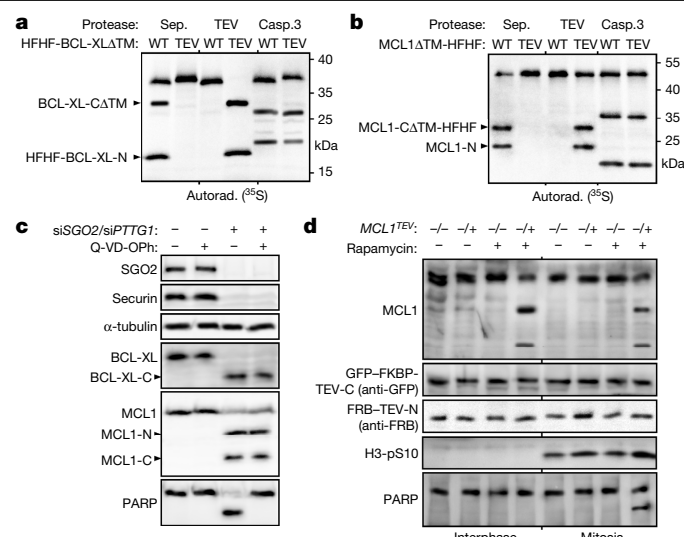


Fig. 1 | The pro-survival factors MCL1 and BCL-XL are separase substrates.

a, b, Autoradiographies of in vitro cleavage assays in the presence of active NEK2A and ATP. WT, wild type; HFHF, His₆-Flag-His₆-Flag tag; ΔTM, transmembrane domain deleted. **c**, siRNA-transfected, taxol-arrested Hek293T cells were treated with Q-VD-Oph or mock-treated and analysed by immunoblotting. siPTTG1, securin depletion. **d**, Immunoblots of parental (–/–) and MCL1^{TEV} heterozygous (–/+) hTERT RPE1 cells after (+) or without (–) rapamycin-induced TEV protease complementation. H3-pS10, Ser10-phosphorylated histone H3 (mitotic marker).

after prolonged mitotic arrest^{14–17}; (2) have been reported to be cleaved by caspases^{18,19}; and (3) contain an ExxR motif that matches the consensus cleavage site of separase close to the caspase cleavage site(s). Indeed, human BCL-XL and MCL1 were cleaved in vitro not only by caspase 3 but also by separase, and the resulting fragments were clearly distinguishable in size (Fig. 1a, b). Replacing 31-ExxR-34 in BCL-XL and 173-ExxR-176 in MCL1 with tobacco etch virus (TEV)-protease cleavage sites required only a few amino acid exchanges (Extended Data Fig. 2a). This rendered both survival factors resistant to separase but susceptible to TEV protease, while leaving cleavage by caspase 3 unaffected (Fig. 1a, b). The proteolytic fragments that were generated by either separase or TEV protease exhibited identical mobilities in SDS-PAGE, thereby indicating that separase cleaves BCL-XL after Arg34 and MCL1 after Arg176. Various in vitro-expressed fragments were used as length standards to confirm the location of the cleavage site for MCL1 (Extended Data Fig. 2b). In mice, the ExxR motif is conserved in BCL-XL but has been replaced with DxxR in MCL1. Still, mouse MCL1 was readily cleaved by separase in vitro (Extended Data Fig. 2c). BCL-XL and MCL1 were also cleaved in human Hek293T, HeLa-K, HCT116, hTERT RPE1 and mouse NIH/3T3 cells during DiM triggered by RNA interference (RNAi) using small interfering RNAs (siRNAs) against *SGO2* and *PTTG1* (which encodes securin) (Extended Data Fig. 2d–g; see also below). Notably, these in vivo cleavages were mediated by separase rather than caspase because (1) Q-VD-Oph, a pan-specific caspase inhibitor, blocked cleavage of PARP but not of BCL-XL and MCL1 (Fig. 1c); (2) fragmentation was not detectable when the endogenous proteins were replaced by their separase-resistant but caspase-sensitive TEV variants (Extended Data Fig. 2d); and (3) MCL1 fragments from Hek293T cells lacking SGO2 and securin perfectly co-migrated with in vitro-expressed fragments comprising amino acids 1–176 (MCL1-N) and 177–350 (MCL1-C) (Extended Data Fig. 2f). In fact, we never observed caspase-dependent processing of MCL1 or BCL-XL in vivo, although this might be due to our focus on early stages of DiM. Using CRISPR–Cas9 gene editing in hTERT RPE1 cells, we replaced the separase cleavage site with a TEV protease cleavage site in one allele of *MCL1*. Upon activation of ‘split TEV protease’ by

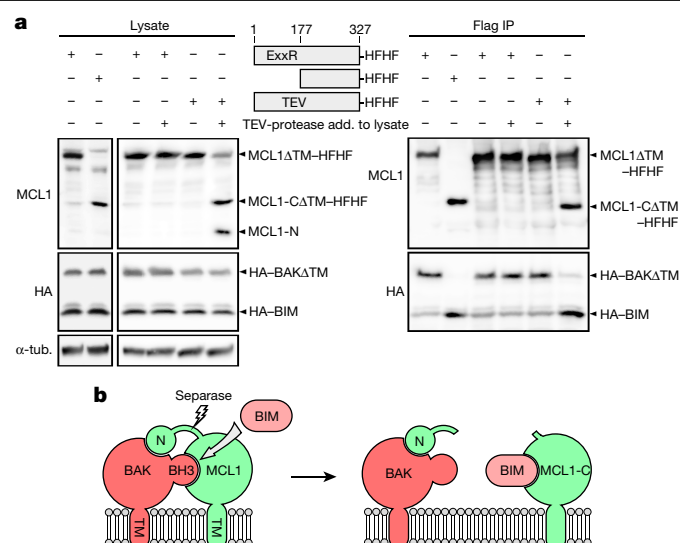


Fig. 2 | Cleavage of MCL1 by separase enables BH3-only proteins to liberate BAK.

a, Immunoblots of (TEV protease supplemented, +) lysates and Flag immunoprecipitation of transfected, mitotic Hek293T cells expressing the indicated, C-terminally His₆-Flag-His₆-Flag (HFHF)-tagged MCL1 fragments together with HA-tagged BAK and BIM. **b**, Model of how separase inactivates MCL1.

rapamycin-induced complementation, half of endogenous MCL1 was cleaved, as expected (Fig. 1d). This was accompanied by considerable PARP cleavage, but only when the cells were in mitosis. Thus, MCL1 cleavage at position 176 is sufficient to initiate DiM during a prometaphase arrest.

Cleavage of MCL1 and BCL-XL liberates BAK

We investigated whether separase-dependent cleavage affected the interactions of MCL1 and BCL-XL with other BCL2 family members. To this end, we used Flag tags to affinity-purify MCL1 and the corresponding C-terminal separase cleavage fragment from transfected Hek293T cells and analysed them for association with co-expressed BAK relative to the BH3-only protein BIM. Whereas MCL1 preferentially interacted with BAK, as expected, MCL1-C bound only to BIM (Fig. 2a). An analogous experiment was conducted with BCL-XL but using BAD instead of BIM owing to the different binding preference of its C-terminal fragment (Extended Data Fig. 3a). Removal of the N-terminal 34 amino acids switched BCL-XL from binding BAK to binding BAD (Extended Data Fig. 3b). For MCL1, the exchange of binding partners upon cleavage was additionally recapitulated by addition of TEV protease to cell lysate containing the TEV variant of MCL1 (Fig. 2a). The BH4 domain of BCL-2 contributes directly to BAX binding²⁰. Similarly, MCL1-N co-purified with BAK from transfected Hek293T cells (Extended Data Fig. 3c), showing that MCL1 contacts BAK not only via its hydrophobic groove but also via its N-terminal domain. We propose that cleavage by separase abolishes the cooperativity of binding, thereby enabling BH3-only proteins such as BIM to supersede BAK from the C-terminal fragment of MCL1 (Fig. 2b). Liberated BAK would then lead to MOMP.

MCL1-N and MCL1-C actively promote apoptosis

We made the counter-intuitive observation that DiM induced by siRNAs against *SGO2* and *PTTG1* was alleviated by co-depletion of MCL1 (Fig. 3a, b) and almost fully rescued by co-depletion of both MCL1 and BCL-XL (Extended Data Fig. 3d, e). While this showed that these two BCL2 family proteins are the crucial—if not the only—substrates of separase during DiM, it also suggested that the separase cleavage fragments of MCL1

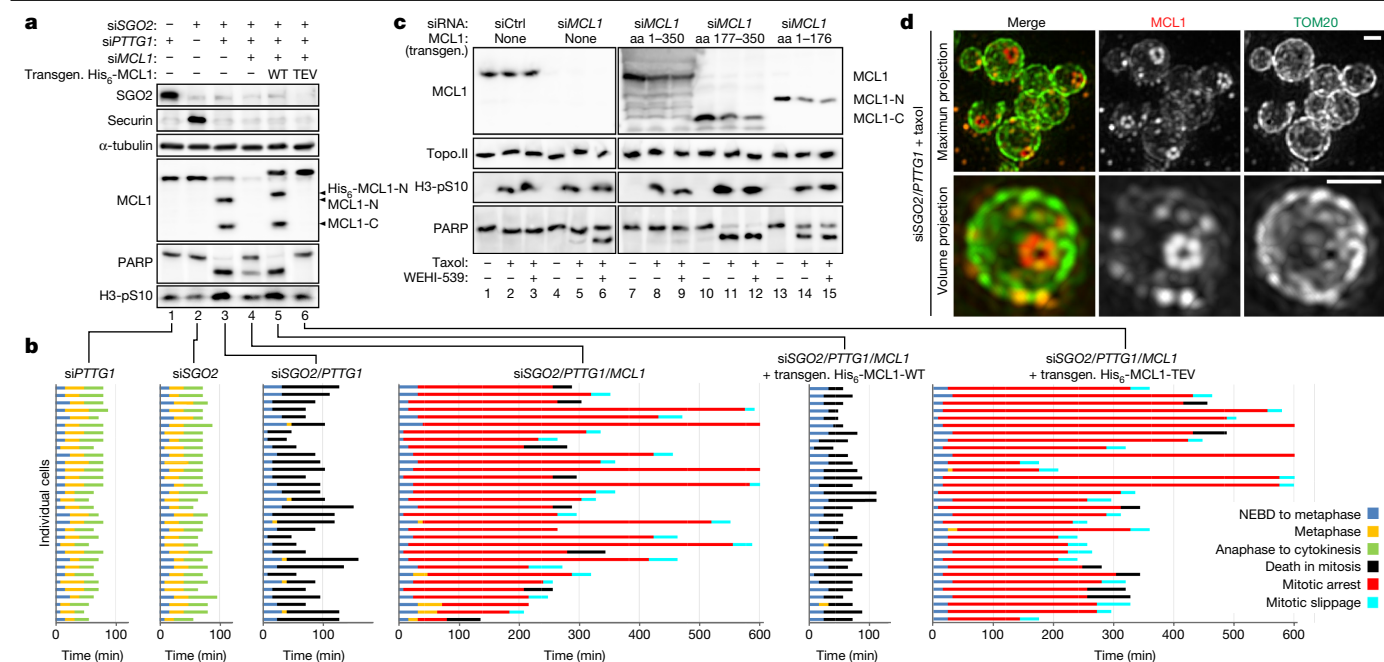


Fig. 3 | Both separase cleavage fragments of MCL1 are pro-apoptotic and MCL1-C forms macropores during late stages of DiM. **a**, **b**, HeLa-K cells transfected with the indicated siRNAs and expression plasmids were released from early S-phase arrest, supplemented with the DNA stain SiR-Hoechst and a fluorogenic caspase-3/7 reporter and analysed by immunoblotting (**a**; 10 h after release) and live cell imaging (**b**; cell fate profiles). **c**, Immunoblots of

MCL1 or control-depleted HeLa-K cells expressing transgenic full-length MCL1 or fragments thereof as indicated. **d**, Immunofluorescence 2D SIM of SGO2- and securin-depleted Hek293T cells undergoing DiM. Note the absence of the mitochondrial outer membrane marker TOM20 from the centres of MCL1 rings. Scale bars, 0.5 μ m.

and BCL-XL are apoptogenic. Consistent with this hypothesis, the mitigation of PARP cleavage by depletion of BCL-XL or MCL1 (Extended Data Fig. 2d, lanes 4–6) was reversed by expression of siRNA-resistant transgenic BCL-XL or MCL1 but not their separase-resistant TEV variants (lanes 7–10). Given the stronger effects of MCL1 depletion, we further studied its cleavage fragments individually. Full-length MCL1, MCL1-N or MCL1-C was expressed in Hek293T cells in which endogenous MCL1 (but not SGO2 or securin) was depleted by RNAi. Short-term absence of MCL1 caused very little apoptosis on its own but led to some cell death in conjunction with chemical inhibition of BCL-XL by WEHI-539 (Fig. 3c, lanes 5 and 6). This phenotype was rescued by transgenic full-length MCL1 (lane 9). Notably, expression of MCL1-N or MCL1-C resulted in PARP cleavage even in the absence of WEHI-539, with the C-terminal fragment having the stronger effect (lanes 11 and 14). PARP cleavage correlated with annexin V and propidium iodide staining as additional markers for apoptosis (Extended Data Fig. 3f, g). Induction of apoptosis by MCL1-N or MCL1-C did not occur in interphase but only in taxol-treated cultures (Fig. 3c, Extended Data Fig. 3f). Time-lapse microscopy further illustrated that HeLa-K cells expressing MCL1-N or MCL1-C also underwent DiM in the absence of spindle toxin (Extended Data Fig. 3h, i). Thus, separase-dependent cleavage not only extinguishes the pro-survival activity of MCL1 but also creates two fragments, each of which kills cells upon entry into mitosis without requiring prolonged mitotic arrest. Investigation of MCL1-N-induced apoptosis suggested that it promotes both liberation of BAK and separase-dependent cleavage of BCL-XL by a positive feedback mechanism (Extended Data Fig. 4).

MCL1-C kills by mitosis-specific MOMP

Interaction analyses of epitope-tagged forms of MCL1 revealed that MCL1-C exhibited homotypic interactions, whereas uncleaved, full-length MCL1 could associate neither with itself nor with MCL1-C (Extended Data Fig. 5a). Self-interaction of MCL1-C and its ability to

induce PARP cleavage were blocked by deletion of the transmembrane domain or presence of the BH3-mimicking MCL1-inhibitor A-1210477, which blocked not only binding of BAK to full-length MCL1 but also binding of BIM to MCL1-C (Extended Data Fig. 5b–d). Tandem affinity purification of co-expressed, differently tagged MCL1-C further revealed that homo-oligomerization and BIM binding are mutually exclusive (Extended Data Fig. 5d), which suggests that BIM interacts with MCL1-C only transiently and disengages upon MCL1-C self-interaction. These observations are reminiscent of the requirements for BAK/BAX-dependent MOMP^{21–23} and, thus, are consistent with pore formation by MCL1-C. Notably, homotypic MCL1-C interactions occurred only in extracts from mitotic, and not interphase, cells (Extended Data Fig. 5b); this is consistent with the finding that the pro-apoptotic effect of MCL1-C is cell-cycle-dependent. According to our hypothesis, separase-induced DiM should be delayed or diminished, but still occur, in the absence of BAK and BAX, whereas MCL1-C-induced DiM should be independent of them. Using time-resolved fractionation of chromatin and organelles, including mitochondria, from the cytosol followed by immunoblotting analyses, we compared parental HCT116 and *BAK1*^{-/-} *BAX*^{-/-} double-knockout cells as they went synchronously from G2- through M-phase. Whereas the kinetics of MCL1-cleavage were indistinguishable in the absence of SGO2 and securin, the release of cytochrome c into the cytosol and PARP cleavage were delayed, but still occurred, in the absence of BAK and BAX (Extended Data Fig. 6a, b). In fact, PARP cleavage and accumulation of Ser139-phosphorylated histone H2A-X (γ H2AX) in the absence of BAK and BAX were less affected during DiM induced by siRNAs against *SGO2* and *PTTG1* than during staurosporine-induced apoptosis²⁴ (Extended Data Fig. 6c). Notably, the timing and extent of cytochrome c release and PARP cleavage were the same in both cell lines when DiM was induced by MCL1-C expression (Extended Data Fig. 6d).

At later stages of intrinsic apoptosis, the mitochondrial network fragments into globular structures, and BAK and BAX form large rings and macropores within the outer membrane^{25–27}. Immunofluorescence

microscopy of MCL1 and the MOM protein TOM20 in taxol-arrested Hek293T cells showed that MCL1 and TOM20 colocalized during DiM, and this colocalization increased when the network dissolved into spheres (Extended Data Fig. 7a, b). Notably, 2D structural illumination microscopy (SIM) revealed the formation of large, typically 0.3- μ m rings by MCL1-C in SGO2- and securin-depleted cells undergoing DiM (Fig. 3d, Extended Data Fig. 7c). TOM20 appeared largely absent from the centre of the rings, suggesting that they represent macropores.

Phosphorylation might explain why the MOMP activity of MCL1-C is specific to mitosis. A candidate approach identified Thr301. Changing this residue to phosphorylation-mimicking Glu enabled MCL1-C to induce apoptosis also in interphase, whereas changing Thr301 to Ala abrogated the pro-apoptotic and self-interaction properties of MCL1-C (Extended Data Figs. 5d, 7d, e). The detection of affinity-purified MCL1-C by a phospho-Thr-specific antibody was limited to mitosis and abolished by the Thr301Ala mutation, suggesting that this position is phosphorylated in vivo (Extended Data Fig. 7d)—possibly by aurora B (Extended Data Fig. 7f, g). Existing structural information places Thr301 of MCL1 at the end of α -helix 6, and dimerization of the corresponding helix has been reported to be involved in homo-oligomerization of BAK^{28,29}. Sequence alignment implies that the Thr301-equivalent position is occupied by a Glu in BAX (Extended Data Fig. 7h). When we changed this constitutively negatively charged residue at position 146 to Ala, the pro-apoptotic function of BAX was abrogated, whereas changing it to Thr rendered BAX a largely mitosis-specific effector of cell death (Extended Data Fig. 7i). A salt bridge at the end of α -helix 6 might therefore be required for pore formation (Extended Data Fig. 7j) and explain why MCL1-C becomes pro-apoptotic only upon phosphorylation of Thr301 during mitosis, whereas MOMP by wild-type BAX is independent of the cell cycle.

Importance of phosphorylation by NEK2A

In unperturbed cells, MCL1 and BCL-XL are present at the onset of anaphase. The question arises of why then cells do not die when separase becomes active on schedule. Considering that cleavage of meiotic kleisin by separase requires its phosphorylation³⁰, we tested the effect of various kinases when first establishing MCL1 and BCL-XL cleavage by separase in vitro (Extended Data Fig. 8). These analyses revealed that cleavage of MCL1 by separase essentially requires NEK2A, while cleavage of BCL-XL was enabled by NEK2A and (less so) CDK1/2–cyclin A2 (note that NEK2A was included in Fig. 1a, b and Extended Data Figs. 2b, c, 4c). NEK2A and CDK1/2–cyclin A are special among mitotic kinases and APC/C substrates in that they are degraded early—that is, at a time when separase activation is still blocked by SAC signalling^{31,32}.

To identify phosphorylation sites within MCL1 and BCL-XL that are relevant for cleavage, we changed candidate serine and threonine residues to alanine or phosphorylation-mimicking acidic residues and screened corresponding variants in kinase and/or cleavage assays. These analyses revealed the following (Extended Data Fig. 9): (1) the NEK2A-dependent phosphorylation of Ser60 and Thr163 is essential for separase-dependent cleavage of MCL1, and phosphorylation of Ser159 further improves it. (2) NEK2A and CDK1/2–cyclin A2 phosphorylate Ser4 and Ser164, and Ser62, respectively, of BCL-XL to enable its cleavage by separase. (3) The phosphorylation-mimicking variants MCL1(S/T60,159,163D/Q) and BCL-XL(S4,62,164D) are cleaved by separase in the absence of kinases.

The above findings suggested that separase does not trigger DiM at anaphase onset merely because NEK2A (and cyclin A2) is absent by then and MCL1 and BCL-XL—owing to dephosphorylation—no longer represent separase substrates (Fig. 4a, top). As a corollary, any temporal overlap between the enzymatic activities of NEK2A and separase should cause DiM. Indeed, this explains why constitutive activity of separase causes early mitotic cell death and why knockdown of NEK2A largely suppressed DiM induced by siRNAs against SGO2 and

PTTG1 (Extended Data Fig. 1g, h). Extending the window of NEK2A activity until separase activation in anaphase should also cause DiM (Fig. 4a, middle). Live cell imaging of transfected HeLa-K cells revealed that overexpression of wild-type NEK2A was compatible with normal mitosis, whereas production of a C-terminally truncated, stabilized variant (Δ MR)³¹ triggered DiM, typically shortly after anaphase onset (Extended Data Figs. 3i, 10a). This was confirmed by time-resolved immunoprecipitation and western analysis of cells synchronously undergoing late mitosis. NEK2A- Δ MR-expressing and mock-treated populations both degraded cyclin B1 and securin and lost SGO2 from separase with similar kinetics (Extended Data Fig. 10b). However, only in NEK2A- Δ MR-containing cells did activation of separase coincide with cleavage of MCL1 and PARP. Consistent with MCL1 and BCL-XL being the relevant targets, apoptosis in anaphase was also triggered by expression of the constitutive separase substrates BCL-XL(S4,62,164D) or MCL1(S/T-60,159,163D/Q) instead of NEK2A- Δ MR (Extended Data Fig. 10c, d). Thus, NEK2A must be degraded in early mitosis to prevent separase from killing cells in anaphase.

Separase-induced DiM increases with MCL1

To investigate whether DiM in response to stabilization of NEK2A was graded with MCL1 dosage, we transfected siRNA or plasmids into Hek293T cells to express wild-type NEK2A or the Δ MR variant and, simultaneously, to reduce or increase the amount of MCL1. When transgenic NEK2A was wild-type and, hence, degraded upon entry into mitosis, or when NEK2A- Δ MR expression was combined with MCL1 depletion, cells showed no signs of apoptosis (Fig. 4b, Extended Data Fig. 11a–c). As seen before, NEK2A- Δ MR induced some DiM at endogenous levels of MCL1. However, annexin V staining and PARP cleavage increased with increasing levels of (transgene-encoded) MCL1 and cleavage fragments thereof. These data suggest that pharmacological inhibition of early mitotic NEK2A degradation should preferentially kill MCL1-overexpressing cells, which are a hallmark of many cancers³³.

A minimal duration of mitosis checkpoint

The SAC is active in each M-phase and chiefly determines its duration¹⁰. Abrogation of the SAC results in chromosomal instability owing to accelerated progression through mitosis². We investigated whether, under these conditions, separase might become active when NEK2A has not yet been fully degraded (Fig. 4a, bottom). When MAD2 and the SAC kinase BUBR1 were depleted by RNAi, HeLa-K cells that had been released from a thymidine arrest degraded securin and cyclin B1 earlier than control cells; this correlated with earlier auto-cleavage of separase, as expected (Fig. 4c). At the same time, degradation of NEK2A and the disappearance of a corresponding MCL1-S60 phosphorylation mark were delayed, which we attribute to competition by other substrates for the APC/C (Fig. 4c, Extended Data Fig. 11d). Notably, this was accompanied by cleavage of MCL1, BCL-XL and PARP and appearance of a sub-G1 peak in flow cytometry, which is another hallmark of apoptotic cells. As seen before, these phenotypes were largely suppressed by co-depletion of MCL1 and BCL-XL and fully suppressed by additional expression of separase-resistant TEV variants of MCL1 and BCL-XL, but re-installed by transfection with the corresponding wild-type transgenes. Cleavage of MCL1, BCL-XL and PARP also occurred upon individual depletion of MAD2 or BUBR1, albeit to lesser extent (Extended Data Fig. 11e). It also occurred in both mouse and human cells upon chemical inhibition of the SAC kinase MPS1 with reversine (Fig. 4d, Extended Data Fig. 11e). Thus, in mammalian cells SAC abrogation suffices to induce cleavage of MCL1 and BCL-XL by separase and consequent DiM. A corollary is that the few existing SAC-deficient tumour cell lines (unless slowed in mitotic progression

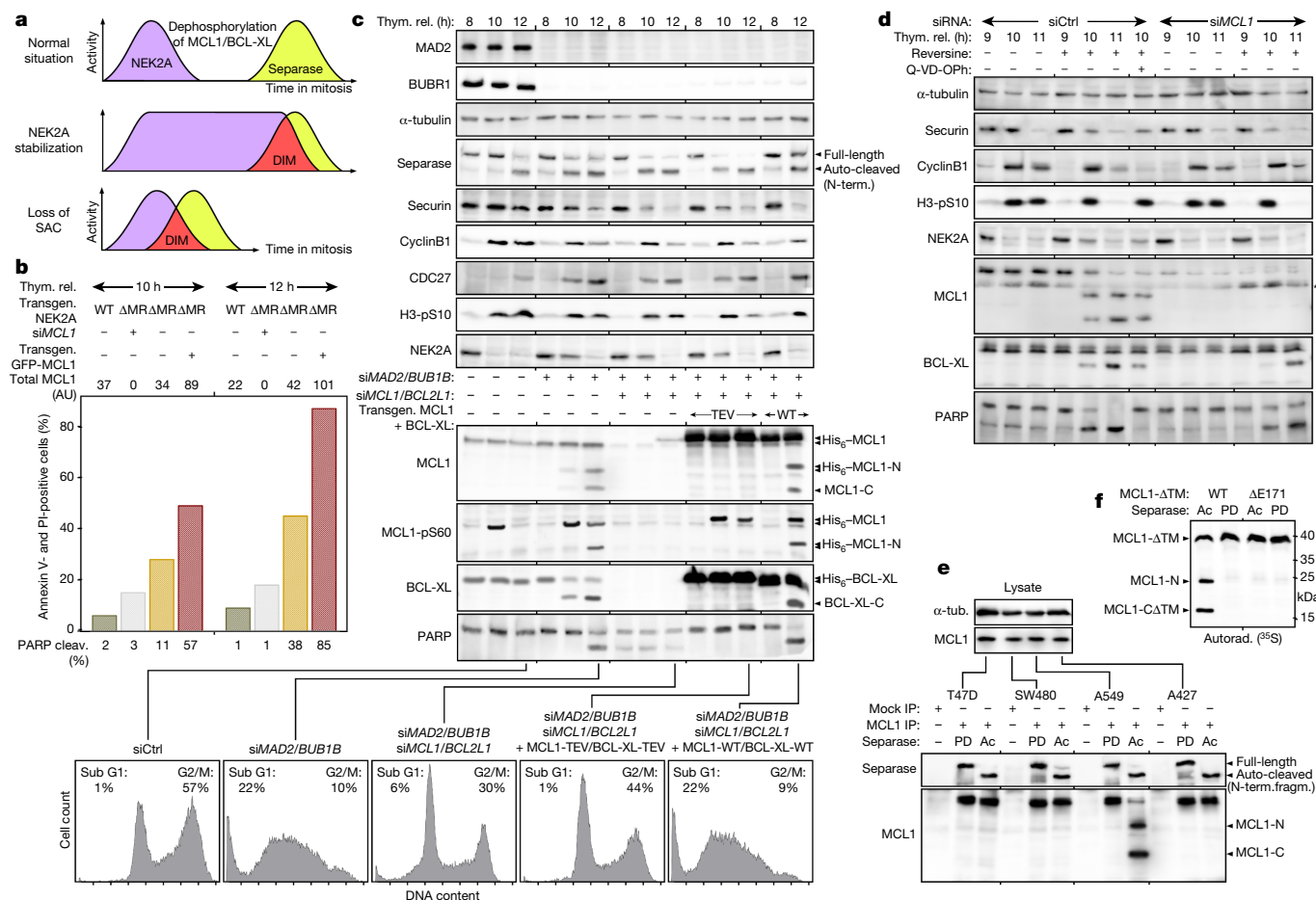


Fig. 4 | DiM due to simultaneous activity of NEK2A and separase is graded by MCL1 level and triggered by SAC deficiency. **a**, Both prolonged activity of NEK2A (middle) and speeded-up mitosis due to SAC deficiency (bottom) cause DiM. **b**, Fluorescence microscopic quantification of annexin V- and propidium iodide (PI)-positive Hek293T cells transfected with siMCL1 and expression plasmids for GFP-MCL1 and NEK2A-WT or NEK2A-ΔMR, as indicated, and released for 10–12 h from thymidine arrest (thy. rel.). Expression of MCL1 and degree of PARP cleavage were quantified by densitometry of immunoblots (Extended Data Fig. 11b). **c**, HeLa-K cells transfected with the indicated siRNAs

and expression plasmids were released from thymidine arrest and analysed by immunoblotting and propidium iodide staining with flow cytometry 8–12 h thereafter. siBUB1A, BUBR1 depletion; siBCL2L1, BCL-XL depletion. **d**, Time-resolved immunoblots of NIH/3T3 cells transfected with indicated siRNAs, pre-synchronized with thymidine, and treated with Q-VD-OPH and/or reversine, undergoing mitosis. Asterisk, nonspecific band. **e**, Immunoblots of lysates and NEK2A/separase-treated Mcl1 immunoprecipitations from the indicated cell lines. **f**, Autoradiograph of in vitro-expressed, NEK2A/separase-treated ³⁵S-MCL1-WT and ³⁵S-MCL1-ΔE171.

by SAC-independent means³⁴) should have found a way to avoid cleavage of MCL1 and BCL-XL. We immunoprecipitated MCL1 from four cancer cell lines: partially SAC-compromised SW480, SAC deficient T47D and A427 and, as a control, SAC-proficient A549 cells^{11,35,36}. Notably, only MCL1 from A549 cells was cleaved upon incubation with NEK2A and separase (Fig. 4e). It is unclear how MCL1 from the other three cell lines is rendered separase-resistant, but it is not due to the mere absence of p53 (Extended Data Fig. 11f). Cancer-associated MCL1 variants seem to be rare. However, of the few catalogued in the Catalogue of Somatic Mutations in Cancer (COSMIC) database³⁷, deletion of Glu171 is by far the most abundant one, being identified ten times in six studies. Although it leaves the ExxR motif intact, unexpectedly, this mutation renders MCL1 resistant to separase (Fig. 4f).

Conclusion

In most cases studied, the intrinsic pathway of apoptosis is triggered by transcriptional upregulation of BH3-only protein expression^{5,7}. Here, we describe a mechanism of DiM, which is probably conserved in vertebrates (Extended Data Fig. 12) and consists of separase-dependent

cleavage of MCL1 and BCL-XL and their concurrent transformation from pro-survival into mitosis-specific pro-apoptotic factors. Our results strongly suggest that the deadliest fragment, MCL1-C, permeabilizes the mitochondrial outer membrane by forming pores. Because separase-dependent cleavage of BCL-XL is also sufficient for DiM, the same might be true of BCL-XL-C. Degradation of MCL1 was causally linked to apoptosis upon mitotic arrest, but exactly how MCL1 is removed remains unknown^{15,16}. Although we have not done so here, it will therefore be interesting to investigate whether MCL1 somehow becomes phosphorylated and separase-activated during prolonged mitosis. Here we studied shortened mitosis. Our results show that abrogation of SAC results in DiM owing to simultaneous activity of NEK2A and separase (Fig. 4a, bottom). We propose that NEK2A and separase form a surveillance mechanism that eliminates SAC-deficient cells that would otherwise be doomed to massive chromosomal instability and aneuploidy, thereby ensuring the survival of cells with the correct length of M-phase and protecting the organism from tumorigenesis. This ‘minimal duration of early mitotic checkpoint’ (DMC) might explain why mutational inactivation of SAC genes in cancer is rare^{35,38}.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-020-2187-y>.

1. Gascoigne, K. E. & Taylor, S. S. Cancer cells display profound intra- and interline variation following prolonged exposure to antimetabolic drugs. *Cancer Cell* **14**, 111–122 (2008).
2. Michel, L. S. et al. MAD2 haplo-insufficiency causes premature anaphase and chromosome instability in mammalian cells. *Nature* **409**, 355–359 (2001).
3. Lin, Z., Luo, X. & Yu, H. Structural basis of cohesin cleavage by separase. *Nature* **532**, 131–134 (2016).
4. Wirth, K. G. et al. Separase: a universal trigger for sister chromatid disjunction but not chromosome cycle progression. *J. Cell Biol.* **172**, 847–860 (2006).
5. Galluzzi, L. et al. Molecular mechanisms of cell death: recommendations of the Nomenclature Committee on Cell Death 2018. *Cell Death Differ.* **25**, 486–541 (2018).
6. Letai, A. et al. Distinct BH3 domains either sensitize or activate mitochondrial apoptosis, serving as prototype cancer therapeutics. *Cancer Cell* **2**, 183–192 (2002).
7. Villunger, A. et al. p53- and drug-induced apoptotic responses mediated by BH3-only proteins puma and noxa. *Science* **302**, 1036–1038 (2003).
8. Kamenz, J. & Hauf, S. Time to split up: dynamics of chromosome separation. *Trends Cell Biol.* **27**, 42–54 (2017).
9. Hellmuth, S. et al. Securin-independent regulation of separase by checkpoint-induced shugoshin–MAD2. *Nature* <https://www.doi.org/10.1038/s41586-020-2182-3> (2020).
10. Taylor, S. S. & McKeon, F. Kinetochore localization of murine Bub1 is required for normal mitotic timing and checkpoint response to spindle damage. *Cell* **89**, 727–735 (1997).
11. Li, Y. & Benezra, R. Identification of a human mitotic checkpoint gene: hSMAD2. *Science* **274**, 246–248 (1996).
12. Rankin, S., Ayad, N. G. & Kirschner, M. W. Sororin, a substrate of the anaphase-promoting complex, is required for sister chromatid cohesion in vertebrates. *Mol. Cell* **18**, 185–200 (2005).
13. Tang, Z., Sun, Y., Harley, S. E., Zou, H. & Yu, H. Human Bub1 protects centromeric sister-chromatid cohesion through Shugoshin during mitosis. *Proc. Natl Acad. Sci. USA* **101**, 18012–18017 (2004).
14. Bennett, A. et al. Inhibition of Bcl-XL sensitizes cells to mitotic blockers, but not mitotic drivers. *Open Biol.* **6**, 160134 (2016).
15. Haschka, M. D. et al. The NOXA-MCL1-BIM axis defines lifespan on extended mitotic arrest. *Nat. Commun.* **6**, 6891 (2015).
16. Sloss, O., Topham, C., Diez, M. & Taylor, S. Mcl-1 dynamics influence mitotic slippage and death in mitosis. *Oncotarget* **7**, 5176–5192 (2016).
17. Topham, C. et al. MYC is a major determinant of mitotic cell fate. *Cancer Cell* **28**, 129–140 (2015).
18. Clem, R. J. et al. Modulation of cell death by Bcl-XL through caspase interaction. *Proc. Natl Acad. Sci. USA* **95**, 554–559 (1998).
19. Michels, J. et al. Mcl-1 is required for Akata6 B-lymphoma cell survival and is converted to a cell death molecule by efficient caspase-mediated cleavage. *Oncogene* **23**, 4818–4827 (2004).
20. Barclay, L. A. et al. Inhibition of pro-apoptotic BAX by a noncanonical interaction mechanism. *Mol. Cell* **57**, 873–886 (2015).
21. Brouwer, J. M. et al. Conversion of Bim-BH3 from activator to inhibitor of Bak through structure-based design. *Mol. Cell* **68**, 659–672.e659 (2017).
22. Czabotar, P. E. et al. Bax crystal structures reveal how BH3 domains activate Bax and nucleate its oligomerization to induce apoptosis. *Cell* **152**, 519–531 (2013).
23. Dai, H. et al. Transient binding of an activator BH3 domain to the Bak BH3-binding groove initiates Bak oligomerization. *J. Cell Biol.* **194**, 39–48 (2011).
24. Wang, C. & Youle, R. J. Predominant requirement of Bax for apoptosis in HCT116 cells is determined by Mcl-1's inhibitory effect on Bak. *Oncogene* **31**, 3177–3189 (2012).
25. Große, L. et al. Bax assembles into large ring-like structures remodeling the mitochondrial outer membrane in apoptosis. *EMBO J.* **35**, 402–413 (2016).
26. Salvador-Gallego, R. et al. Bax assembly into rings and arcs in apoptotic mitochondria is linked to membrane pores. *EMBO J.* **35**, 389–401 (2016).
27. McArthur, K. et al. BAK/BAX macropores facilitate mitochondrial herniation and mtDNA efflux during apoptosis. *Science* **359**, eaao6047 (2018).
28. Day, C. L. et al. Solution structure of pro-survival Mcl-1 and characterization of its binding by proapoptotic BH3-only ligands. *J. Biol. Chem.* **280**, 4738–4744 (2005).
29. Dewson, G. et al. Bak activation for apoptosis involves oligomerization of dimers via their $\alpha 6$ helices. *Mol. Cell* **36**, 696–703 (2009).
30. Kudo, N. R. et al. Role of cleavage by separase of the Rec8 kleisin subunit of cohesin during mammalian meiosis I. *J. Cell Sci.* **122**, 2686–2698 (2009).
31. Hayes, M. J. et al. Early mitotic degradation of Nek2A depends on Cdc20-independent interaction with the APC/C. *Nat. Cell Biol.* **8**, 607–614 (2006).
32. Wolthuis, R. et al. Cdc20 and Cks direct the spindle checkpoint-independent destruction of cyclin A. *Mol. Cell* **30**, 290–302 (2008).
33. Beroukhim, R. et al. The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899–905 (2010).
34. Wild, T. et al. The spindle assembly checkpoint is not essential for viability of human cells with genetically lowered APC/C activity. *Cell Rep.* **14**, 1829–1840 (2016).
35. Tighe, A., Johnson, V. L., Albertella, M. & Taylor, S. S. Aneuploid colon cancer cells have a robust spindle checkpoint. *EMBO Rep.* **2**, 609–614 (2001).
36. Weitzel, D. H. & Vandr , D. D. Differential spindle assembly checkpoint response in human lung adenocarcinoma cells. *Cell Tissue Res.* **300**, 57–65 (2000).
37. Bamford, S. et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br. J. Cancer* **91**, 355–358 (2004).
38. Hernando, E. et al. Molecular analyses of the mitotic checkpoint components hSMAD2, hBUB1 and hBUB3 in human cancer. *Int. J. Cancer* **95**, 223–227 (2001).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

  The Author(s), under exclusive licence to Springer Nature Limited 2020

Article

Methods

Antibodies

Antibodies generated within this study were raised and purified as described⁹. The following antibodies were used for immunoblotting and immunofluorescence microscopy (IFM): Rabbit anti-separase³⁹, rabbit anti-sororin⁴⁰, rabbit anti-PIN1⁴¹, goat anti-CDC27⁴², mouse anti-GFP⁴³, mouse anti-Flag (1:2,000; Sigma-Aldrich, M2), rabbit anti-SGO2 (1:1,000; Bethyl, A301-262A), rabbit or guinea pig anti-SGO2 (1.5 µg/ml; anti-DVPPRESHSDQSSKC), rabbit anti-SGO1 (1:500; Abcam, ab21633), mouse anti-securin (1:1,000; MBL, DCS-280), rabbit anti-phosphoSer10-histone H3 ('H3-pS10'; 1:1,000; Millipore, 06-570), mouse anti-cyclin B1 (1:1,000; Millipore, 05-373), rabbit anti-cleaved caspase 3 (Asp175; 1:1,000; Cell Signaling, 5A1E), rabbit anti-BCL-XL (1:1,000; Cell Signaling, 2762), mouse anti-MCL1 (1:800; BioLegend, W16014A), guinea-pig anti-MCL1 (0.75 µg/ml; for IFM 1.5 µg/ml; raised against amino acids 1–327 (ΔTM) of human MCL1), guinea-pig anti-phosphoSer10-MCL1 ('MCL1-pS60'; 0.5 µg/ml; for IFM 1 µg/ml; anti-CVIGGpSAGA, liberated from reactivity towards CVIGGSAGA), anti-TOM20 (1:500; Santa-Cruz Biotechnology, F10), mouse anti-cytochrome c (1:1,000; BD Pharmingen, 7H8.2C12), mouse anti-BubR1 (1:1,000; BD Transduction Laboratories, clone 9), rabbit anti-BAX (1:1,000; Abcam, ab32503), rabbit anti-BAK (1:1,000; Abcam, ab32371), rabbit anti-MAD2 (1:1,000; Bethyl, A300-300A), rabbit anti-PARP (1:800; Cell Signaling, 46D11), mouse anti-PARP (1:1,000; Calbiochem, AM30), anti-MBP monoclonal (1:1,000; NEB Biolabs, E8038S, HRP-conjugated), mouse anti-NEK2 (1:600; BD Transduction Laboratories, clone 20), mouse anti-RGS-His₆ (1:1,000; Qiagen 34610), rabbit anti-phosphoSer139-histone H2A.X ('γH2AX'; 1:5,000; Millipore, EP854(2)Y), mouse anti-topoisomerase IIα (1:1,000; Enzo Life Sciences, IC5), mouse anti-cyclin A2 (1:200; Santa Cruz Biotechnology, 46B11), rat anti-HA (1:2,000; Roche, 3F10) and mouse anti-α-tubulin (hybridoma supernatant 1:200; DSHB, 12G10). Nonspecific rabbit, mouse and guinea pig IgGs were from Sigma-Aldrich. For immunoprecipitation experiments, the following affinity matrices and antibodies were used: mouse anti-Flag agarose (Sigma-Aldrich, M2), rat anti-HA agarose (Roche, clone 3F10), anti-GFP nanobody covalently coupled to NHS-agarose (GE Healthcare), mouse anti-RGS-His₆ or guinea pig anti-MCL1 coupled to protein G sepharose (GE Healthcare) and rabbit anti-BCL-XL coupled to protein A sepharose (GE Healthcare). For non-covalent coupling of antibodies, 10 µl of the respective matrix was rotated with 2–5 µg antibody in the presence of 1% w/v BSA (Roth) for 90 min at room temperature and then washed three times. Secondary antibodies for immunoblotting were horseradish peroxidase (HRP)-conjugated goat anti-rabbit, anti-mouse and anti-guinea pig IgGs (Sigma-Aldrich, all used at 1:20,000). The following secondary antibodies were used for IFM (all 1:500): Cy3 donkey anti-guinea pig IgG, Cy5 goat anti-mouse IgG (both Jackson ImmunoResearch Laboratories), Alexa Fluor 488 goat anti-mouse IgG (Invitrogen). Antisera against mouse SGO2 and mouse securin were gifts from A. M. Pendás.

Cell lines and plasmids

HeLa-K, Hek293T, hTERT RPE1, HCT116, HCT116 *BAK*^{-/-}, *BAX*^{-/-}, HCT116 *TP53*^{-/-}, SW480, and NIH/3T3 cells were cultured in Dulbecco's modified Eagle's medium (DMEM) (Biowest), T-47D cells in RPMI-1640 (Biowest) and A427 and A549 cells in essential minimum Eagle's medium (EMEM) (Roth). All media were supplemented with 10% heat-inactivated fetal calf serum (FCS) (Sigma-Aldrich). Cells were cultured at 37 °C in 5% CO₂. *Xenopus laevis* S3 cells were grown at 27 °C under atmospheric CO₂ in 70% Leibovitz's L-15 medium (Gibco) supplemented with 1% Glutamax (Gibco) and 10% heat-inactivated FCS. *MCL1* (NM_021960) was PCR-cloned from human testis cDNA (Clontech), *BCL-XL* (*BCL2L1*, NM_138578), *BAK1* (*BCL2L7*, NM_001188), *BIM* (*BCL2L11*, NM_001204106) and *BAD* (*BCL2L8*, NM_032989) were PCR-cloned from self-made HeLa cell cDNA. *MCL1* from *X. laevis* (NP_001131055), *Xenopus tropicalis*

(XP_002935512) and *Danio rerio* (NP_571674) were PCR-cloned from self-made oocyte cDNA, zygote cDNA (a gift from C. Niehrs) and 72-h embryos (a gift from P. Braaker), respectively. Details about the resulting plasmids and derivatives thereof are available upon request. The ZipGFP plasmid was a gift from X. Shu (Addgene plasmid #81241)⁴⁴. For transient expression of proteins, Hek293T, HeLa-K or hTERT RPE1 cells were transfected with corresponding pCS2- or pcDNA5-based plasmids using a calcium phosphate-based method or Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. *X. laevis* S3 cells were transfected using PEI (Polysciences; 3 µl of 1 µg/µl per 1 µg DNA). *MCL1*^{+/TEV} hTERT RPE1 cells were generated using the co-CRISPR approach⁴⁵ with 0.8 µM ouabain octahydrate (Sigma) 72 h after transfection with 30 pmol ssODN *MCL1*^{TEV} (5'-GCTGGAGTTGGTCGGGGAATC TGGAATAACACCAGTACGGACGGGTCACTACCCTCGACGCCGCCGCC AGCAGAGGAGGAGGAGGACGAGAAGTGTACTTCCAGTCGCTCGAGA TTATCTCTCGGTACCTTCGGGAGC-3'), 10 pmol ssODN *ATPIA1* and 1 µg eSpCas9. *ATPIA1_G3_Dual_sgRNA* (a gift from Y. Doyon, Addgene plasmid #86613) encoding the corresponding sgRNA for *MCL1*^{TEV} (5'-CGAGTTGTACCGGCAGTCGC-3'). Ouabain-resistant clones were screened by PCR with oligonucleotides (fwd: 5'-GAGTTC GCTGGCGCCACCCCGTAGGACT-3', rev: 5'-GGGAGTGAGCCTTGG CGATTAATGAACCCCTT-3') and the resulting 871-bp fragment was further analysed for the presence of a XhoI site.

Cell treatments

For synchronization in early S-phase, human and frog cells were treated with 2 mM thymidine (Sigma-Aldrich) for 20 h. Upon release/wash-out, cells entered mitosis within 8–10 h. Synchronization of cells in prometaphase was done by addition of taxol (LC Laboratories) to 0.2 µg/ml 6 h after release from thymidine block. G2 arrest was achieved by addition of 10 µM RO-3306 (Santa-Cruz Biotechnology) 4 h after release from thymidine arrest for 6–10 h. For release from G2 arrest, cells were trypsinized, washed 5× with fresh medium and reseeded for the indicated incubation times. To inhibit MCL1, BCL-XL, caspases or MPS1 kinase, cells were supplemented with A-1210477 (2.5 µM, Abcam), WEHI-539 (0.5 µM, Cayman Chemicals), Q-VD-OPh (20 µM, BD Pharmingen) or reversine (1 or 5 µM, Cayman Chemicals), respectively, at the time of taxol addition and incubated for 6 h. To assess apoptosis in interphase, cells were arrested with thymidine for 20 h together with simultaneous transfection of corresponding plasmids or siRNA, released for 15 h, re-supplemented with thymidine and analysed 10 h thereafter. To assess apoptosis in mitosis, cells were transfected 10–12 h before thymidine addition, released and treated with taxol as described above, and collected when morphological signs of apoptosis first became visible (typically 10–13 h after thymidine wash-out). When studying depletion of MCL1 and BCL-XL without separate deregulation, the taxol arrest was prolonged to 12 h before cells were analysed. As control, cells were treated with 1 µM staurosporine (Abcam) for 8–12 h. To address DiM in SAC-abrogated HeLa-K cells, the corresponding plasmids were transfected first. Ten hours later, siRNA was transfected. Thymidine was added 14 h thereafter. (In the case of NIH/3T3 cells, thymidine (4 mM) was added immediately after siRNA transfection and washed away 18 h thereafter. This was followed by a second thymidine block 10 h later.) After 20 h, cells were released, re-seeded, supplemented with reversine (5 µM unless stated otherwise) and incubated for the indicated times without addition of taxol. For the 'taxol-ZM override' experiments, taxol-arrested HeLa-K cells were collected by shake-off and released for the indicated times by replating into medium supplemented with ZM447439 (5 µM, Tocris Biosciences), taxol (0.2 µg/ml) and cycloheximide (30 µg/ml, Sigma-Aldrich). To induce FRB-FKBP heterodimerization of the split TEV⁴⁶, rapamycin (100 nM, Sigma) was added 10 h before cells were collected.

Immunoprecipitation and subcellular fractionation

Cells (1 × 10⁷) were lysed with a dounce homogenizer in 1 ml LP2 lysis buffer (20 mM Tris-HCl (pH 7.7), 100 mM NaCl, 10 mM NaF, 20 mM

β -glycerophosphate, 5 mM $MgCl_2$, 0.1% Triton X-100, 5% glycerol), supplemented with benzonase (30 U/l; Santa Cruz) and complete protease inhibitor cocktail (Roche), and incubated on ice for 1 h. To preserve phosphorylations, lysis buffer was additionally supplemented with calyculin A (50 nM, LC-Laboratories) and microcystin LR (1 μ M, Alexis Biochemicals). If transmembrane proteins were to be analysed, the corresponding lysis reactions were cleared by low-speed centrifugation (2,500g for 10 min), giving rise to whole-cell extracts (WCE). In all other cases, lysis reactions were cleared by centrifugation at 16,000g for 30 min, resulting in lysates. For immunoprecipitations, 1 ml of WCE or lysate was rotated over 10 μ l of antibody-carrying beads for 4–12 h at 4 °C and washed 5 \times with lysis buffer. For Extended Data Figs. 4c and 8c, immobilized MCL1 or BCL-XL–BAK/–BAD complexes were incubated with separase before boiling in SDS-sample buffer. For Fig. 4e, incubation also included NEK2A and ATP. For cleavage of MCL1-TEV in lysate (Fig. 2a), 20 U of His₆-TEV protease was added at 18 °C and immunoprecipitation was started 30 min thereafter. Intact mitochondria were enriched and separated from cytosol as described²⁷.

RNA interference

For efficient knockdown, cells were transfected with calcium phosphate or RNAiMax (Invitrogen) using 70–100 nM siRNA duplex directed against *PTTG1* (*SECURIN*): 5'-UCUUAGUGCUUCAGAGUUUGUGUGUAU-3', *SGO2*: 5'-GAACACAUUUUCUUCGCCUATT-3', *ESPL1* (*SEPARASE*): 5'-AACUGUUCUACCUCCAAGGUUAAGAUUU-3', *NEK2A*: 3'-UUCUGAGAGUCAGCUCACA-5', *MCL1*: 5'-CGAAGGAAGUAUCGAUUUUTT-3', *BCL2L1* (*BCL-XL*): 5'-CCAGGGAGCUUGAAAGUUUUTT-3', *SGO1*: 5'-GAUGACAGCUCCAGAAAUUTT-3', *CDC45* (*SORORIN*): 3'-UGGAGGAGCUCGAGACGGA-5', *BUB1B* (*BUBR1*): 5'-GGACACATTTAGATGCACTTT-3' and/or *MAD2*: 5'-GCTTGTAAGTACTGATCTTTT-3'. For transfer of siRNA into NIH/3T3 cells, we used Lipofectamine 2000 (Invitrogen) according to an optimized protocol supplied by the manufacturer. The following pre-designed siRNAs (IDT) were used: mm.RI.MCL1.13.1: 5'-GAGUGCUGACUAGAUGAUAACUUAUUAUCUAGUC-3', mm.RI.MCL1.13.2: 5'-GCGUAAACCAAGAAAGCUUCGAUGAAGCUUUCUUGG-3' (a mixture of both was used for Fig. 4d), mm.RI.PTTG1.13.1 ('a' in Extended Data Fig. 2g): 5'-UAUCUUUGUUGAUAGGAUUAUUAUCCUUAUUAAC-3', mm.RI.PTTG1.13.2 ('b' in Extended Data Fig. 2g): 5'-AUCACCGAGAAGUCUACUGUGUCUUAGUAGACUUCU-3', mm.RI.SGOL2a.13.1 ('a' in Extended Data Fig. 2g): 5'-ACCUCUUCAGUAUCAAGAAAGGUUGUCUUGAUACUG-3' and mm.RI.SGOL2a.13.2 ('b' in Extended Data Fig. 2g): 5'-GAAACUUAGACAAAAGGUUCGAUUUACUUUUGUCU-3'. *Luciferase* siRNA (*GL2*) served as control (Ctrl). If depletion of the corresponding protein caused DiM, siRNA transfection was performed during thymidine block and cells were collected upon entering mitosis. In all other cases transfection was performed on asynchronous cells 12 h before synchronization.

Fluorescence microscopy

To detect apoptosis in fixed samples, Hek293T cells were grown on poly-lysine-coated glass coverslips and processed 10–12 h after release from thymidine arrest. Different staining and fixation procedures were carried out. Where indicated, cells were stained with annexin V–FITC and propidium iodide according to the manufacturer's protocol (Annexin V–FITC Apoptosis Detection Kit, Abcam), washed once with corresponding binding buffer and mounted onto coverslips in 5 μ l DAPI-Fix (1 \times MMR, 48% glycerol, 11% formaldehyde, 1 mg/ml Hoechst 33342)⁴⁷. To label intact mitochondria, 200 nM MitoTracker (Orange CMTMRos, Invitrogen) was added 10 h after release from thymidine to Hek293T cells in serum-free cell culture medium and incubated for 45 min before cells were fixed. If additional antibody staining was performed, corresponding coverslips were washed once with PBS, fixed with fixation solution (PBS, 3.7% formaldehyde) for 15 min at room temperature, and then washed twice with quenching solution (PBS, 100 mM glycine). Cells were then treated with permeabilization solution

(PBS, 0.5% Triton X-100) for 5 min, washed with PBS and incubated in blocking solution (PBS, 1% (w/v) BSA) overnight at 4 °C. Coverslips were transferred into a wet chamber, incubated with primary antibodies for 1 h, washed four times with PBS-Tx (PBS, 0.1% Triton X-100), incubated with fluorescently labelled secondary antibodies for 1 h, washed once with PBS-Tx, stained for 10 min with 1 μ g/ml Hoechst 33342 in PBS-Tx, washed four times and mounted in 20 mM Tris-HCl (pH 8.0), 2.33% (w/v) 1,4-diazabicyclo(2.2.2)octane, 78% glycerol on a glass slide. Immunofluorescence microscopy of fixed cells was performed on a DMI 6000 inverted microscope (Leica) using a HCX PL APO 100 \times /1.40–0.70 oil objective. Z-stack series were collected in 0.2- μ m increments over 10 μ m, deconvoluted (blind algorithm) and, where indicated, projected into one plane using the LAS-AF software. For 2D SIM, cells were grown on Precision cover glasses (Marienfeld) and imaged with the Nikon Eclipse Ti2 using a SR APO TIRF AC 100 \times H objective. Z-stacks were captured in 0.1- μ m increments over 2 μ m, processed using the stack reconstruction mode and visualized by volume and maximum projection of the NIS-Elements AR software (Nikon). Chromosome spreads were prepared as described⁴⁸. For video microscopy, transfected HeLa-K or *X. laevis* S3 cells were released from single 20 h thymidine block (how long?), transferred into a μ -Slide 8-well (Ibidi) dish, and imaged starting 5–6 h later over a period of 5–15 h in 8–10-min intervals on a DMI 6000 inverted microscope (Leica) using a HCX PL APO 40 \times /0.85 CORR (HeLa) or HCX PL FLUOTAR L 20 \times /0.40 CORR PH1 (S3) objective and the corresponding LAS AF600 software. To visualize chromatin, SiR-Hoechst (200 nM)⁴⁹ and Verapamil (1 μ M, both from Spirochrome) were added to the culture medium 6 h before imaging. Where indicated, medium was additionally supplemented with IncuCyte Caspase 3/7 reagent (5 μ M, Essen Bioscience)⁵⁰.

In vitro kinase and cleavage assay

MCL1 from different origins served as substrates for in vitro phosphorylation reactions. Immunoprecipitated MCL1 bound to 10 μ l beads, 3 μ l of in vitro translated MCL1 Δ TM (TNT Quick-coupled Transcription and Translation kit, reticulocyte lysate-based, Promega) or 2 μ g bacterially expressed, purified MCL1 Δ TM were combined with 'cold' ATP (5 μ M) and 40 μ Ci γ -³²P-ATP (Hartmann Analytic) for radioactive labelling or only with cold ATP (1 mM) for non-radioactive phosphorylation. Reactions of 25 μ l were assembled in kinase buffer (10 mM Hepes-KOH (pH 7.7), 50 mM NaCl, 25 mM NaF, 1 mM EGTA, 20% glycerol, 10 mM $MgCl_2$, 10 mM DTT) including 1 μ l of NEK2A- Δ MR, NEK2B- Δ MR or Δ 86-cyclinA2–CDK1/2 (corresponding to 12.5 \times 10⁵ transfected Hek293T cells each; see below), 0.4 μ g of His₆-tagged Δ 90-cyclinB1–CDK1⁵¹, 0.1 μ g of PLK1 (ProKinase, No. 0183-0000-1) or 0.1 μ g of GST-tagged Aurora B (PRECISIO-Kinase, A2108). To specifically inhibit phosphorylation, the following kinase inhibitors were used: RO-3306 (2 μ M, Santa-Cruz), BI-2536 (100 nM, Boehringer-Ingelheim), staurosporine (300 nM, Abcam) and ZM-447439 (0.5 μ M, Tocris). For reactions containing BCL-XL Δ TM, 3 μ l IVT served as substrate. To test kinase activity, 2 μ g each of histone H1 (NEB) or myelin basic protein (MBP, Upstate Biotechnology) were used. Kinase reactions were incubated for 30 min at 37 °C, subjected to SDS–PAGE, blotted onto PVDF membrane (SERVA) and analysed by autoradiography using a phospho-sensitive imaging plate (Fujifilm). The same membrane was re-activated with methanol and further analysed by immunoblotting. For cleavage assays, 12.5- μ l reactions were incubated with 1 μ l separase (active P1127A variant or protease-dead (PD) C2029S variant)^{39,41}, 1 U human caspase 3 (Enzo) or 20 U TEV protease, incubated for 30 min at room temperature (separase) or 37 °C (both others) and stopped by addition of SDS-sample buffer.

Recombinant protein expression and purification

To produce recombinant NEK2A- Δ MR (active or kinase-dead (KD, L37M), NEK2B- Δ MR and Δ 86-cyclinA2 (in complex with endogenous CDK1/2), 10 \times 10⁷ Hek293T cells were transfected with the corresponding plasmids to express the kinases in fusion with a GFP-SUMOstar tag⁵²,

supplemented with taxol 24–48 h thereafter and collected 12 h later. Lysates in LP2 (including protease and phosphatase inhibitors) were cleared by centrifugation for 30 min at 16,500g and rotated for 4 h at 4 °C with anti-GFP nanobody beads (0.1 ml corresponding to 1 mg nanobody). Immobilized kinases were washed three times in LP2 (200g, 1 min, 4 °C), transferred into cleavage buffer (10 mM Hepes-KOH (pH 7.7), 50 mM NaCl, 25 mM NaF, 1 mM EGTA, 20% glycerol) and incubated with 40 µg His₆-SUMOstar protease⁵² for 45 min at 18 °C. Eluates (80 µl) were snap-frozen and stored in aliquots at –80 °C. His₆-SUMO1–MCL1ΔTM was expressed from a pET28-derivative in *Escherichia coli*. Rosetta 2 DE3 (Novagen). Bacteria were lysed in LP1 (PBS, 5 mM imidazole, 0.5 mM DTT and an additional 400 mM NaCl) and purified over Ni²⁺-NTA-agarose (Qiagen) according to standard procedures. The eluate in PBS supplemented with 250 mM imidazole, 0.5 mM DTT and an additional 400 mM NaCl (pH adjusted to 7.5 with HCl) was dialysed against LP1 for 12 h at 4 °C in the presence of 10 ng His₆-Senp2⁵³ per 100 µg MCL1 and then rotated for 3 h over 9/10th the amount of fresh Ni²⁺-NTA-agarose. The flow-through containing pure MCL1ΔTM was dialysed against PBS and stored in aliquots at –80 °C.

Flow cytometry

To measure apoptosis in vivo, Hela-K cells transfected with corresponding siRNAs and plasmids were collected by gentle trypsinization, washed once with medium, and then stained with annexin V-FITC and propidium iodide according to the manufacturer's protocol (Abcam, ab14085). Samples were analysed immediately on a Cytomix FC 500 (Beckman Coulter) using an FL1 signal detector for FITC and FL2 for propidium iodide and counting at least 20,000 single cells per condition. Flow cytometry of propidium iodide-stained cells was done as described⁴².

Statistics and reproducibility

No statistical methods were used to predetermine sample size. Experiments analysed by immunoblotting or autoradiography were repeated 2–4 times with similar results (2–4 biological replicates). For quantitative analyses of chromosome spreads, clonogenic assays, and IFM specimens, the investigators were blinded to sample allocation. Otherwise, experiments were not randomized.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data that support the findings of this study are available within the paper. Source Data for Figs. 1–4 and Extended Data Figs. 1–12 are provided with the paper. Data or other materials are available from the corresponding author upon reasonable request.

39. Stemmann, O., Zou, H., Gerber, S. A., Gygi, S. P. & Kirschner, M. W. Dual inhibition of sister chromatid separation at metaphase. *Cell* **107**, 715–726 (2001).
40. Wolf, P. G., Cuba Ramos, A., Kenzel, J., Neumann, B. & Stemmann, O. Studying meiotic cohesin in somatic cells reveals that Rec8-containing cohesin requires Stag3 to function and is regulated by Wapl and sororin. *J. Cell Sci.* **131**, jcs212100 (2018).
41. Hellmuth, S. et al. Human chromosome segregation involves multi-layered regulation of separase by the peptidyl-prolyl-isomerase Pin1. *Mol. Cell* **58**, 495–506 (2015).
42. Hellmuth, S., Böttger, F., Pan, C., Mann, M. & Stemmann, O. PP2A delays APC/C-dependent degradation of separase-associated but not free securin. *EMBO J.* **33**, 1134–1147 (2014).
43. Hellmuth, S., Gutiérrez-Caballero, C., Llano, E., Pendás, A. M. & Stemmann, O. Local activation of mammalian separase in interphase promotes double-strand break repair and prevents oncogenic transformation. *EMBO J.* **37**, e99184 (2018).
44. To, T. L. et al. Rational design of a GFP-based fluorogenic caspase reporter for imaging apoptosis in vivo. *Cell Chem. Biol.* **23**, 875–882 (2016).
45. Agudelo, D. et al. Marker-free coselection for CRISPR-driven genome editing in human cells. *Nat. Methods* **14**, 615–620 (2017).
46. Wehr, M. C. et al. Monitoring regulated protein–protein interactions using split TEV. *Nat. Methods* **3**, 985–993 (2006).
47. Murray, A. W. Cell cycle extracts. *Methods Cell Biol.* **36**, 581–605 (1991).
48. McGuinness, B. E., Hirota, T., Kudo, N. R., Peters, J. M. & Nasmyth, K. Shugoshin prevents dissociation of cohesin from centromeres during mitosis in vertebrate cells. *PLoS Biol.* **3**, e86 (2005).
49. Lukinavičius, G. et al. SiR-Hoechst is a far-red DNA stain for live-cell nanoscopy. *Nat. Commun.* **6**, 8497 (2015).
50. Hanson, K. M. & Finkelstein, J. N. An accessible and high-throughput strategy of continuously monitoring apoptosis by fluorescent detection of caspase activation. *Anal. Biochem.* **564–565**, 96–101 (2019).
51. Gorr, I. H., Boos, D. & Stemmann, O. Mutual inhibition of separase and Cdk1 by two-step complex formation. *Mol. Cell* **19**, 135–141 (2005).
52. Liu, L., Spurrier, J., Butt, T. R. & Strickler, J. E. Enhanced protein expression in the baculovirus/insect cell system using engineered SUMO fusions. *Protein Expr. Purif.* **62**, 21–28 (2008).
53. Butt, T. R., Edavettal, S. C., Hall, J. P. & Mattern, M. R. SUMO fusion technology for difficult-to-express proteins. *Protein Expr. Purif.* **43**, 1–9 (2005).
54. Hames, R. S. & Fry, A. M. Alternative splice variants of the human centrosome kinase Nek2 exhibit distinct patterns of expression in mitosis. *Biochem. J.* **361**, 77–85 (2002).
55. Inoshita, S. et al. Phosphorylation and inactivation of myeloid cell leukemia 1 by JNK in response to oxidative stress. *J. Biol. Chem.* **277**, 43730–43734 (2002).
56. Maurer, U., Charvet, C., Wagman, A. S., Dejardin, E. & Green, D. R. Glycogen synthase kinase-3 regulates mitochondrial outer membrane permeabilization and apoptosis by destabilization of MCL-1. *Mol. Cell* **21**, 749–760 (2006).

Acknowledgements We thank T. U. Mayer for suggesting the concept of the DMC, R. Youle and M. Orth for cell lines, D. Pfeiffer for help with the 2D SIM, S. Heidmann, T. Klecker, and P. Wolf for critical reading of the manuscript, and J. Hübner and M. Hermann for technical assistance. This work was supported by a grant (STE997/4-2) from the Deutsche Forschungsgemeinschaft (DFG) to O.S.

Author contributions S.H. carried out all experiments. S.H. and O.S. co-designed the research and wrote the paper.

Competing interests The authors declare no competing interests.

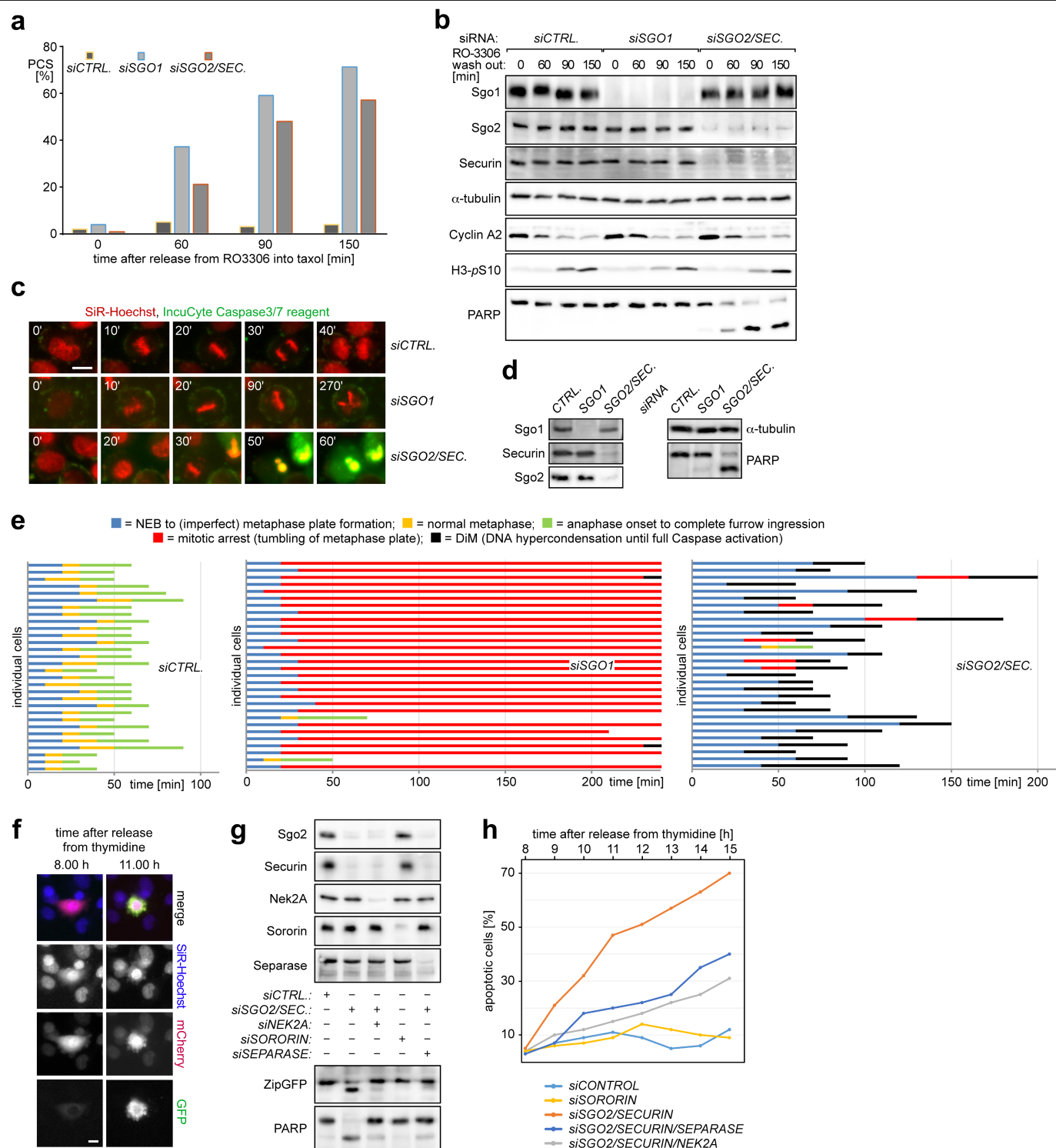
Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41586-020-2187-y>.

Correspondence and requests for materials should be addressed to O.S.

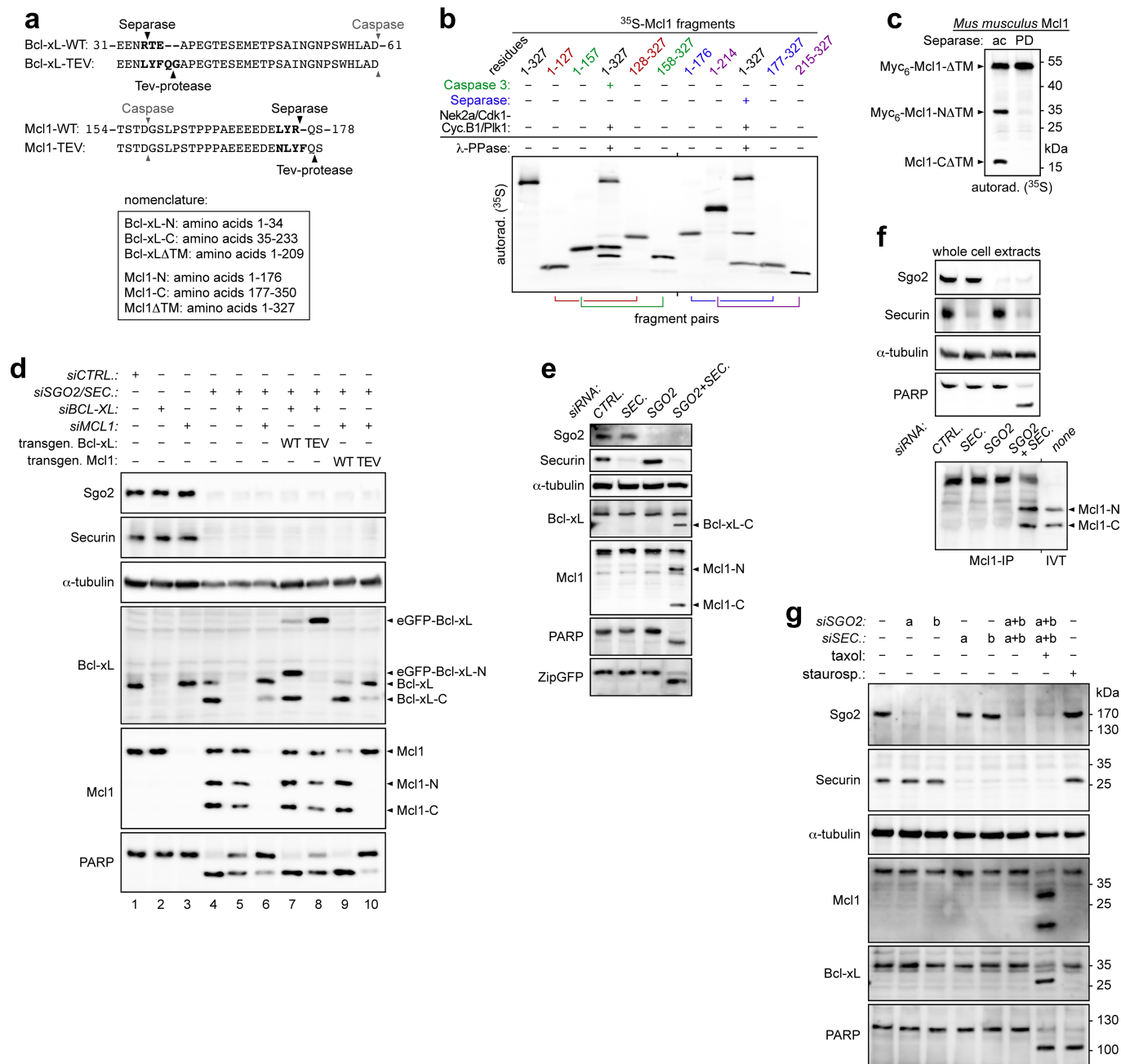
Peer review information Nature thanks Andreas Villunger, Hongtao Yu and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at <http://www.nature.com/reprints>.



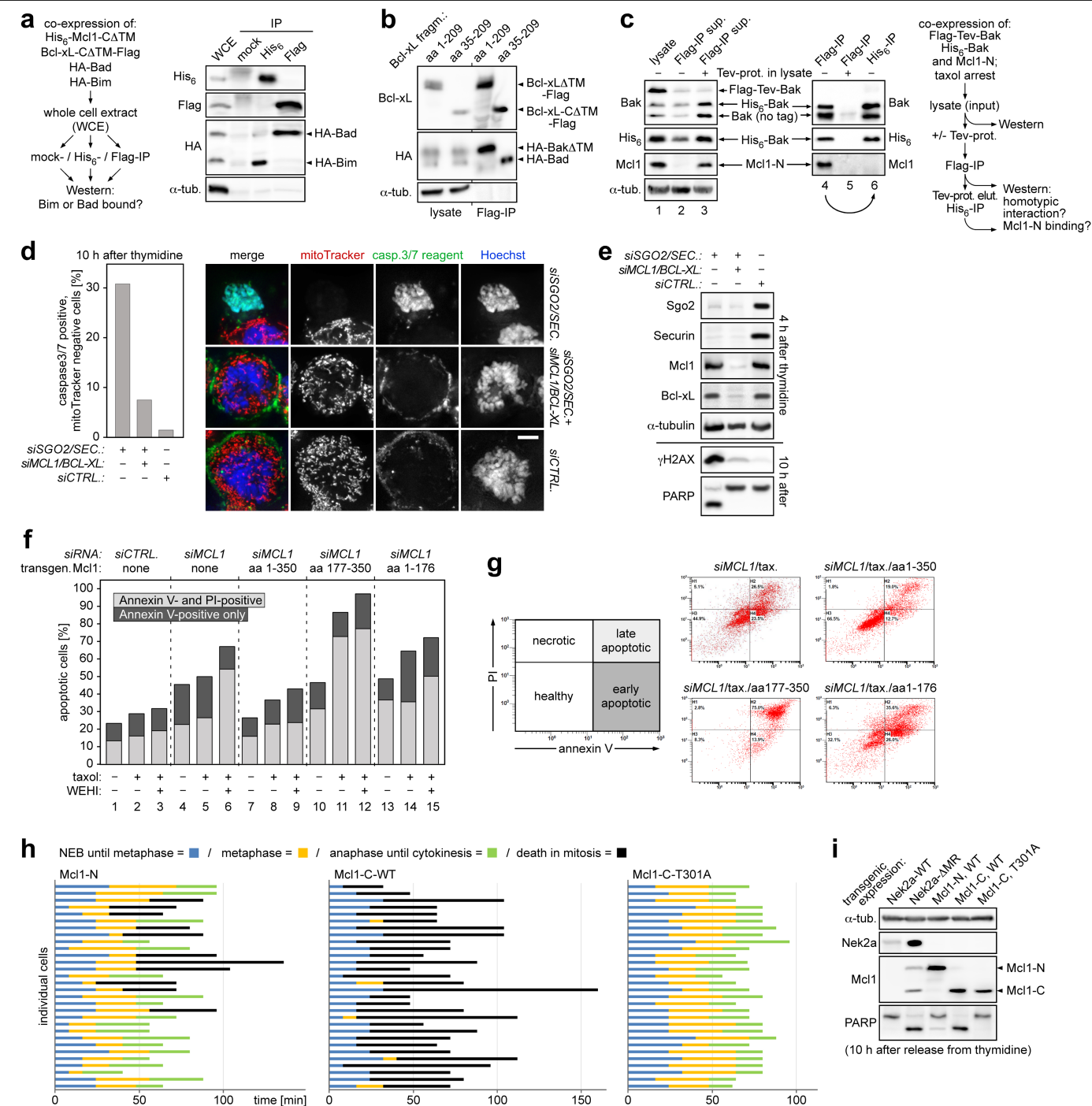
Extended Data Fig. 1 | Premature activation of separase rather than loss of cohesion triggers DiM. **a**, Premature sister chromatid separation (PCS) was quantified by chromosome spreading from siRNA-transfected HeLa-K cells at different times after release from RO-3306/G2 arrest. **b**, Immunoblots of cells from **a**. **c–e**, HeLa-K cells transfected with the indicated siRNAs and cultured in the presence of SiR-Hoechst⁴⁹ and a fluorogenic caspase 3/7 reporter⁵⁰ were analysed by video fluorescence microscopy. Shown are representative stills

(**c**; scale bar, 10 μ m), immunoblots (**d**), and cell fate profiles (**e**). **f–h**, HeLa-K cells expressing the caspase 3 reporter ZipGFP⁴⁴ were transfected with the indicated siRNAs, supplemented with SiR-Hoechst and followed by video microscopy. At least 100 cells were counted per time point and condition. Shown are representative stills (**f**), immunoblots (**g**), and line graphs (**h**) of the percentages of GFP-positive (apoptotic) cells. The ZipGFP plasmid also expresses mCherry as a control. Scale bar, 10 μ m.



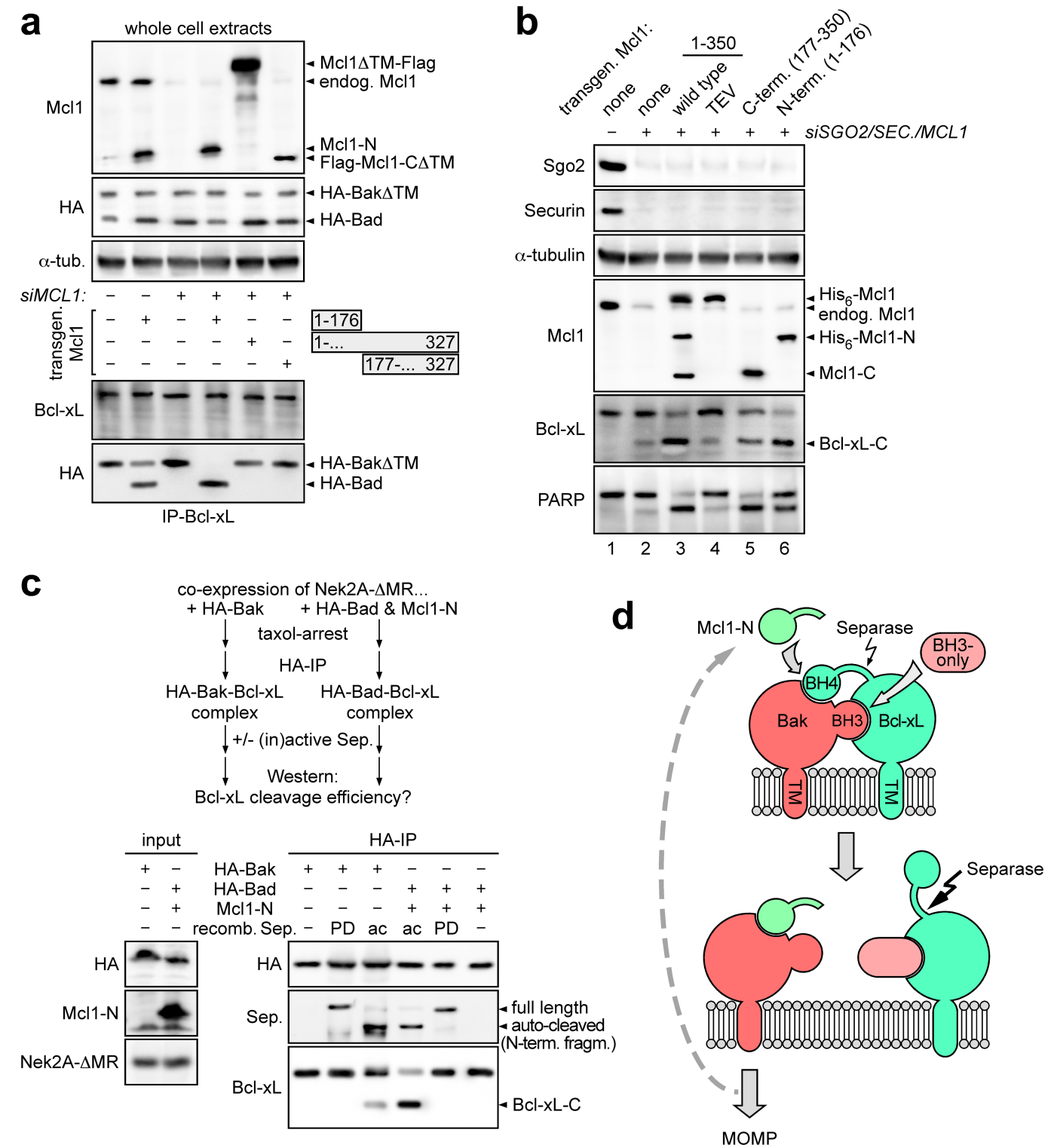
Extended Data Fig. 2 | Characterization of MCL1 (and BCL-XL) cleavage by separase (and caspase 3). **a**, Sequence stretches of wild-type BCL-XL and MCL1 and TEV variants thereof. Differing amino acids are in bold. Arrowheads show protease cleavage sites. **b**, Separase and caspase 3 cleave MCL1 after Arg176 and Asp157, respectively, as mapped by in vitro-expressed fragments. Following incubation with NEK2A, CDK1-cyclin B1, PLK1 and ATP, separase, or caspase 3, in vitro-translated ³⁵S-MCL1ΔTM was treated with a surplus of λ-PPase and analysed by SDS-PAGE and autoradiography. ³⁵S-MCL1 fragments representing reported caspase 3 cleavage fragments and putative separase cleavage fragments served as molecular weight standards. **c**, Mouse MCL1 is cleaved by separase after 154-DXXR-157. Autoradiograph of in vitro-translated, NEK2A/separase-treated mouse ³⁵S-MCL1. PD, protease-dead (C2029S); Ac, active (P1127A); ΔTM, transmembrane domain deleted. **d**, Immunoblots of

taxol-arrested Hek293T cells transfected with siRNAs and expression plasmids as indicated. During separase-triggered DiM, MCL1 cleavage stimulates BCL-XL cleavage and vice versa, which—at least in the case of MCL1—is mediated by the corresponding N-terminal fragment (Extended Data Fig. 4b). **e**, Co-depletion of SGO2 and securin induces cleavage of MCL1 and BCL-XL followed by apoptosis in non-transformed cells. Immunoblots of siRNA-transfected, taxol-arrested hTERT RPE1 cells expressing ZipGFP. **f**, MCL1 is cleaved after R176 during DiM of SGO2- and securin-depleted cells. Immunoblots of extracts and MCL1 immunoprecipitates (IP) from taxol-arrested Hek293T cells transfected with the indicated siRNAs. IVT, in vitro translated. **g**, Mouse MCL1 and BCL-XL are cleaved and DiM is triggered upon separase deregulation in mouse cells. NIH/3T3s cells transfected with siRNAs and treated with taxol and staurosporine as indicated were analysed by immunoblotting.



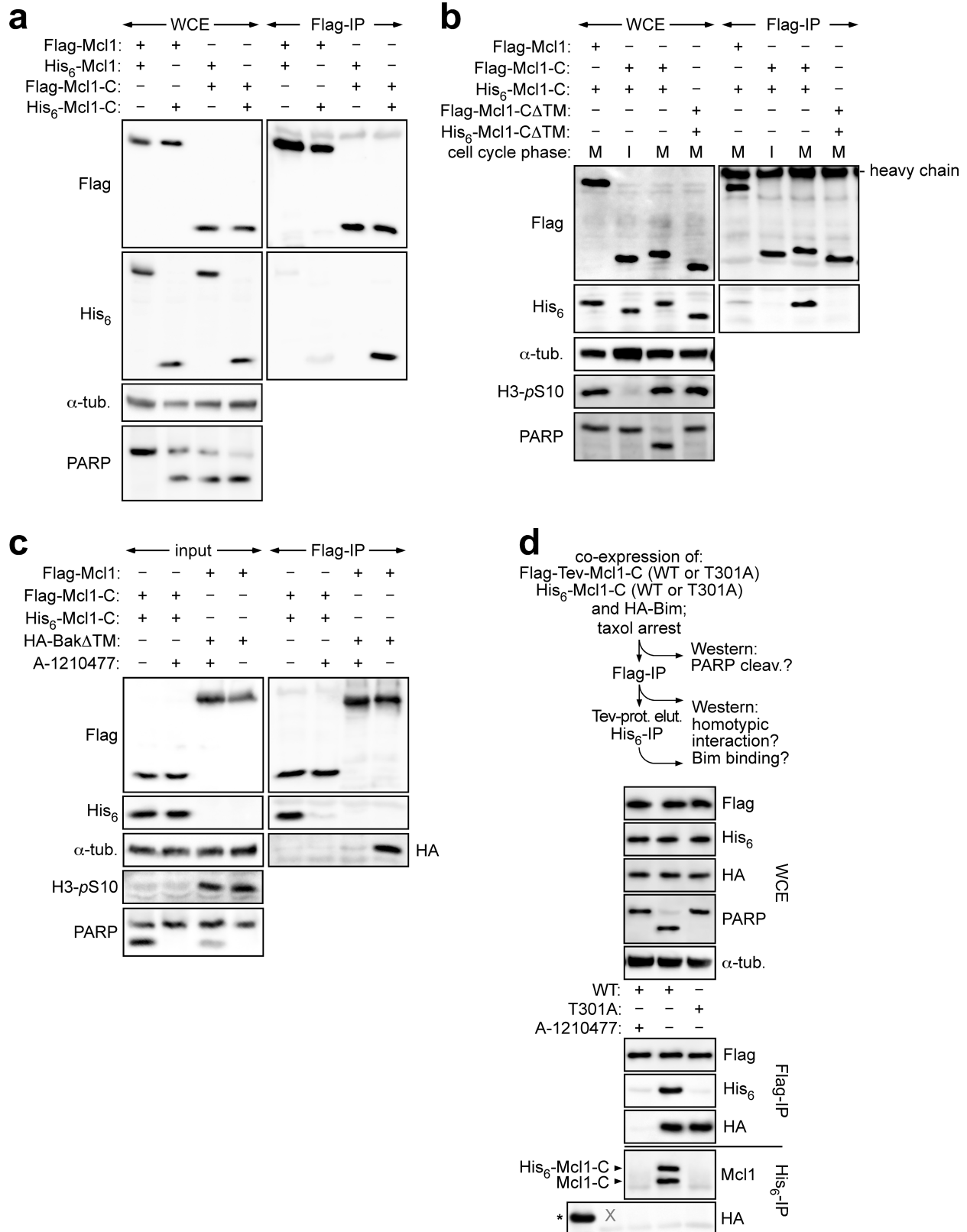
Extended Data Fig. 3 | Characterization of separate cleavage fragments of MCL1 and BCL-XL. **a**, The C-terminal separate cleavage fragments of MCL1 and BCL-XL preferentially bind BIM and BAD, respectively. Experimental setup and immunoblots of the indicated immunoprecipitation from taxol-arrested Hek293T cells co-expressing His₆-MCL1-CATM, BCL-XL-CATM-Flag, HA-BAD, and HA-BIM. **b**, BCL-XL and BCL-XL-C bind BAK and BAD, respectively. Immunoblots of Flag immunoprecipitation from transfected, mitotic Hek293T cells expressing the indicated Flag-tagged BCL-XL fragments together with HA-tagged BAK and BAD. **c**, MCL1-N interacts with BAK. Experimental setup and immunoblots of lysate and consecutive Flag and His₆ immunoprecipitations from transfected, mitotic Hek293T cells co-expressing Flag-TEV-BAK, His₆-BAK, and MCL1-N. TEV protease supplementation of lysate served as a negative control. Self-interaction of BAK is mutually exclusive with binding of MCL1-N. **d**, Separese-induced DiM is suppressed by knock-down of

MCL1 and BCL-XL. Quantification and representative images of siRNA-transfected, mitotic Hek293T cells cultivated in the presence of mitoTracker and a fluorogenic caspase 3/7 reporter before fixation and Hoechst staining. At least 100 cells each were counted. Scale bar, 5 μm. **e**, Immunoblots of cells from **d**. **f**, MCL1-N and -C promote DiM. Plot of early (dark grey) and late (light grey) apoptosis as judged by flow cytometric analysis of propidium iodide and annexin V staining of Hek293T cells transfected with siRNAs and expression vectors for transgenic MCL1 (fragments) and supplemented with taxol and BCL-XL inhibitor WEHI-539 (WEHI) as indicated. **g**, Representative 2D scatter plots of cells from **f** and their interpretation. **h**, Induction of DiM by MCL1-C does not require taxol treatment but does require Thr301. Cell fate profiles of HeLa-K cells expressing the indicated MCL1 fragments and cultured in the presence of SiR-Hoechst and a fluorogenic caspase 3/7 reporter. **i**, Immunoblot of cells from **h** and Extended Data Fig. 10a.



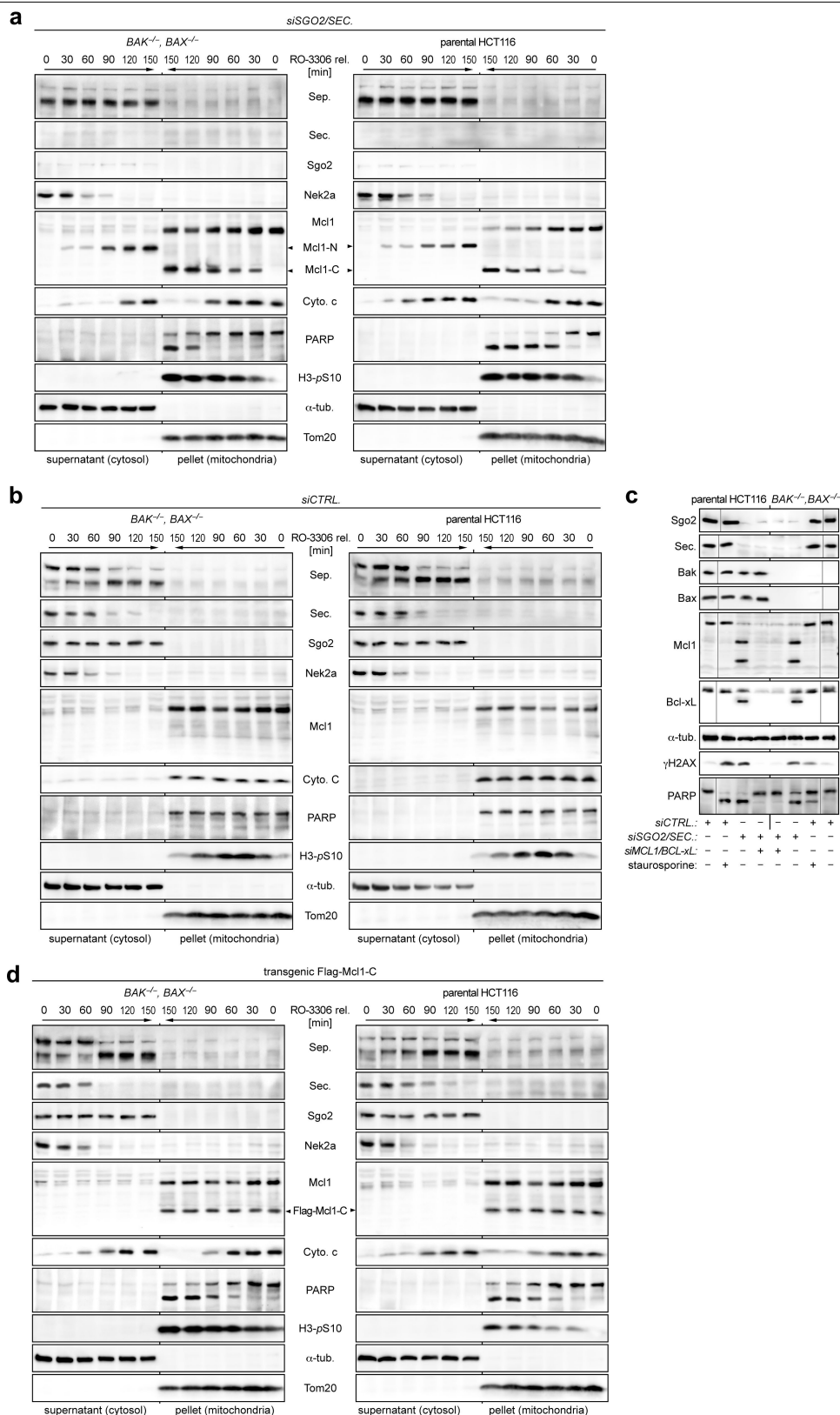
Extended Data Fig. 4 | MCL1-N enables BAD to replace BAK as an interactor of BCL-XL and enhances cleavage of BCL-XL by separase. **a**, MCL1-N causes a switch in the binding partner of BCL-XL from BAK to BAD. Immunoblots of BCL-XL immunoprecipitation from MCL1-depleted or control-treated, mitotic Hek293T cells co-expressing HA-BAK, HA-BAD and, where indicated, various MCL1 fragments. Separase was not deregulated in this experiment, which is why BCL-XL is not cleaved. **b**, Separase-dependent BCL-XL cleavage in cells is primarily stimulated by MCL1-N. Immunoblots of SGO2/securin/MCL1 triple-depleted or control-treated, prometaphase Hek293T cells expressing the indicated transgenic MCL1 variants. **c**, Separase prefers BCL-XL in complex with BAD as a substrate rather than BAK. Experimental setup and immunoblots

of HA immunoprecipitation from taxol-arrested Hek293T cells expressing NEK2A-ΔMR and either HA-BAK or HA-BAD plus MCL1-N. Before SDS-PAGE, samples were incubated with inactive (PD) or active (ac) separase or control treated (-). **d**, MCL1-N promotes apoptosis by a positive feedback mechanism. MCL1-N competitively displaces the BH4 domain of BCL-XL from BAK, resulting in BH3-only proteins, such as BAD, excluding BAK from BCL-XL. At the same time, this renders BCL-XL a better separase substrate. MCL1-N acts catalytically, being released from BAK upon self-interaction and pore formation by the latter (dotted arrow; Extended Data Fig. 3c).



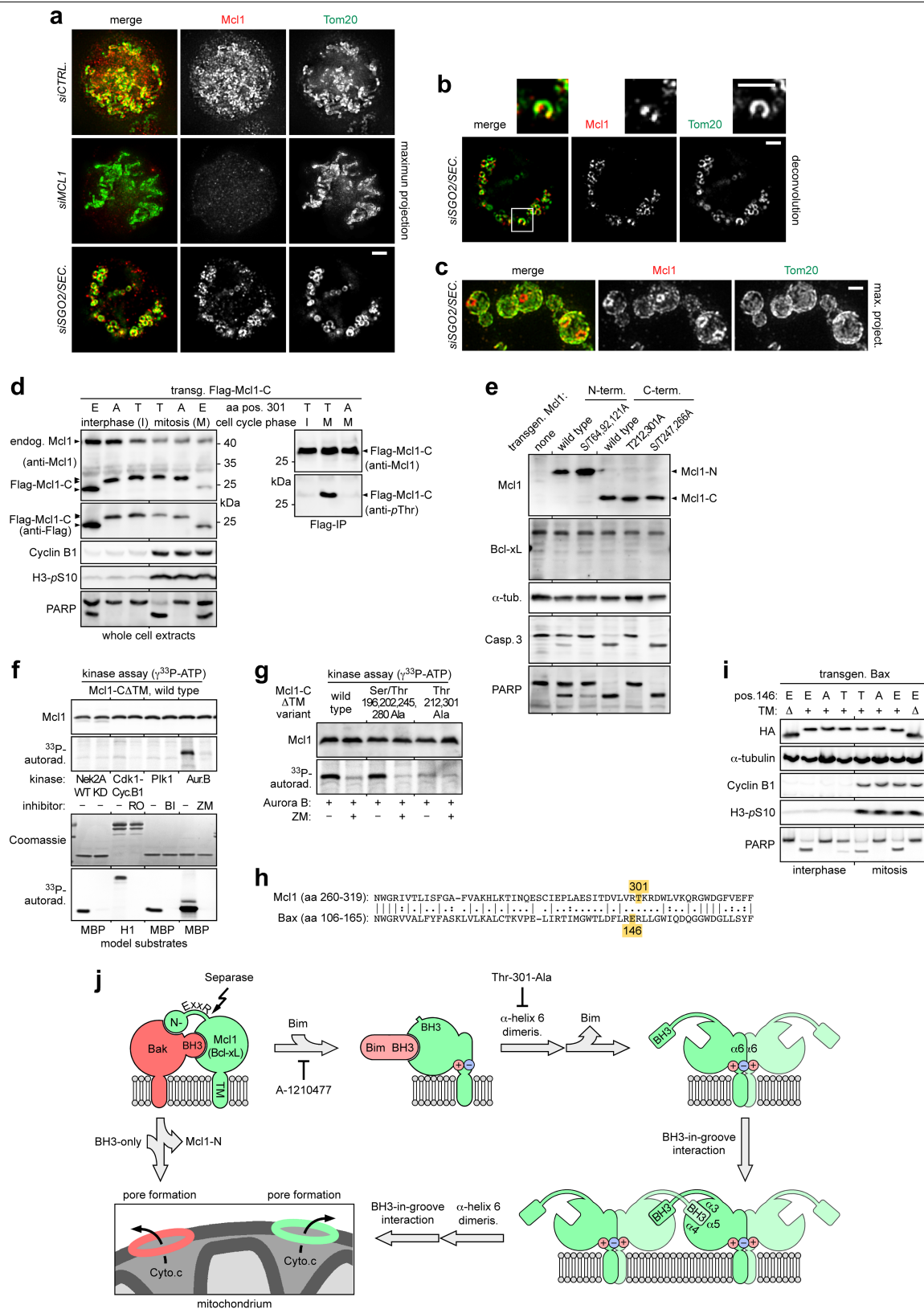
Extended Data Fig. 5 | Self-interaction of pro-apoptotic MCL1-C shares characteristics with pore formation by BAK/BAX but requires mitosis-specific phosphorylation. a–c. MCL1-C exhibits mitosis-specific self-interaction, which requires the transmembrane domain and an accessible BH3-binding groove. Immunoblots of Flag immunoprecipitation from mitotic or interphase Hek293T cells expressing MCL1 variants and supplemented with the MCL1 inhibitor A-1210477 as indicated. Blockade of the MCL1–BAK interaction served as a control for the effectiveness of A-1210477. **d.** The

homotypic interaction of MCL1-C is mutually exclusive with BIM binding. Top, experimental setup; bottom, immunoblots of consecutive Flag and His₆ immunoprecipitations from Hek293T cells expressing HA–BIM together with Flag–TEV- and His₆-tagged forms of either wild-type MCL1-C or the T301A variant. X, irrelevant lane between HA–BIM control (asterisk) and His₆ immunoprecipitation samples. The T301A mutation prevents self-interaction of MCL1-C but not its association with BIM.



Extended Data Fig. 6 | BAK/BAX-independent release of cytochrome c by separase deregulation or MCL1-C expression. **a, b**, Time-resolved immunoblots of cytosol and mitochondria-containing fractions from SGO2/securin-depleted or mock-transfected *BAK^{-/-}, BAX^{-/-}* and parental HCT116 cells released from a G2-arrest at $t = 0$ min. **c**, Immunoblots of taxol-arrested *BAK^{-/-}, BAX^{-/-}* and parental HCT116 cells that were transfected with siRNA and supplemented with staurosporine as indicated. Note the absence of MCL1 and

BCL-XL cleavage during staurosporine-induced apoptosis and suppression of *siSGO2/siPTTG1*-induced DiM by co-depletion of MCL1 and BCL-XL. Grey lines within panels are between lanes that were not directly juxtaposed but nevertheless originated from the same gel. **d**, Time-resolved immunoblots of cytosol and mitochondria-containing fractions from Flag-MCL1-C-expressing *BAK^{-/-}, BAX^{-/-}* and parental HCT116 cells released from G2-arrest at $t = 0$ min.

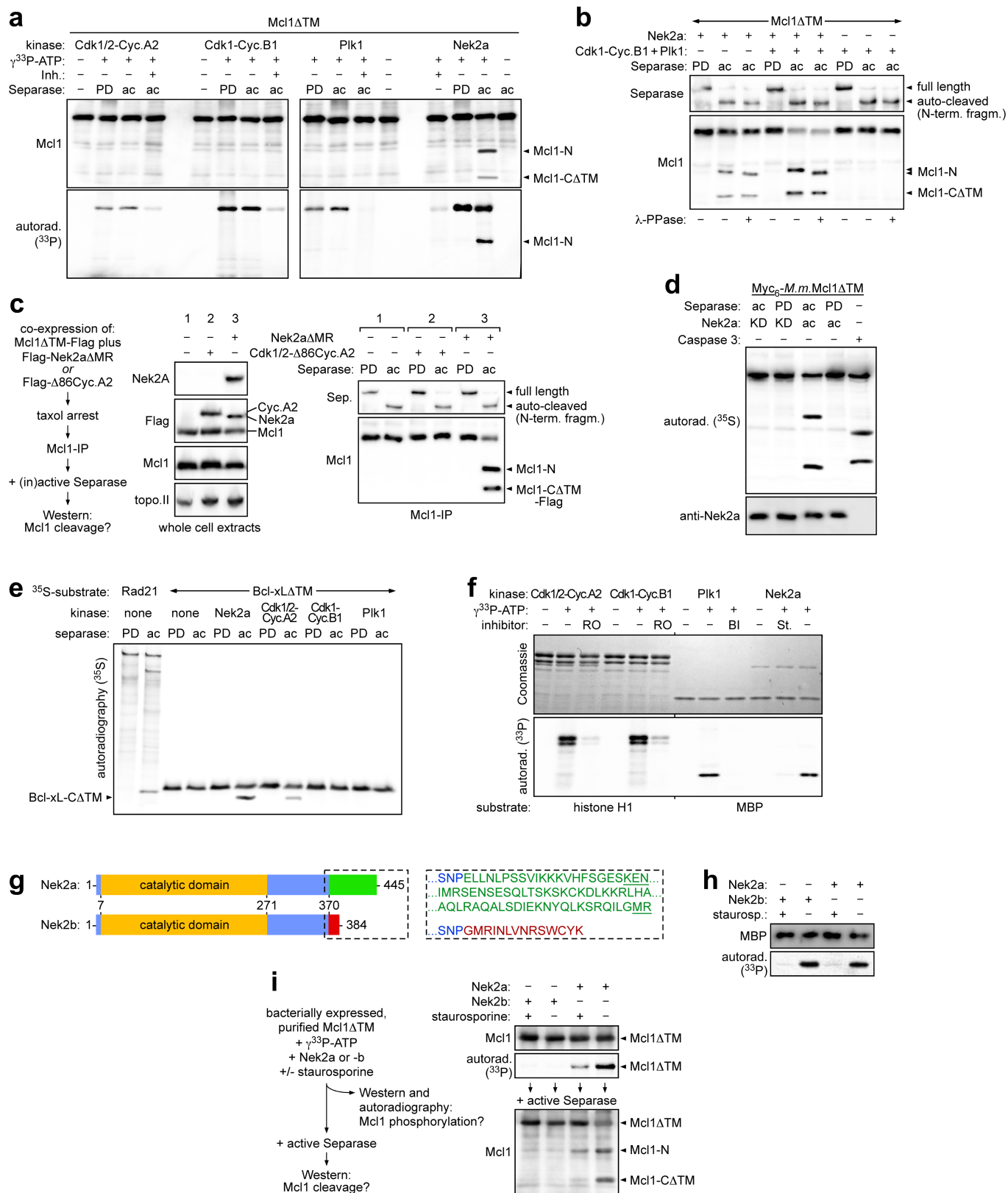


Extended Data Fig. 7 | See next page for caption.

Article

Extended Data Fig. 7 | The pro-apoptotic effect of MCL1-C and BAX requires a negative charge at the end of α -helix 6. **a, b**, Immunofluorescence micrographs of taxol-arrested Hek293T cells transfected with the indicated siRNAs. Shown are maximum projections of 20 z-stacks (**a**) or a single, deconvoluted plane (**b**). The interruption of TOM20 rings by MCL1 dots is consistent with a cross-section through a fragmented mitochondrion containing an MCL1 ring. Scale bars, 3 μ m. **c**, MCL1-C is likely to form macropores into the mitochondrial outer membrane. Immunofluorescence 2D SIM of SGO2/securin-depleted Hek293T cells undergoing DiM. Note the absence of the mitochondrial outer membrane marker TOM20 from the centres of MCL1 rings. Scale bar, 0.5 μ m. **d**, Immunoblots of extracts and Flag immunoprecipitations from interphase or mitotic Hek293T cells expressing Flag-tagged MCL1-C-WT (T), MCL1-C(T301A) (A), or MCL1-C(T301E) (E). **e**, Identifying serine and threonine residues that affect the pro-apoptotic nature of MCL1-N and MCL1-C. Immunoblot of MCL1-depleted, taxol-arrested Hek293T cells expressing the indicated siRNA-resistant variants of MCL1-N or MCL1-C. BCL-XL is not cleaved during apoptosis if separase remains inhibited.

f, g, Aurora B kinase phosphorylates MCL1-C in vitro, probably at position 301 primarily. Immunoblots and autoradiographs of in vitro-translated wild-type MCL1-C Δ TM and variants thereof after incubation with the indicated kinases and inhibitors in the presence of $\gamma^{33}\text{P}$ -ATP. Activity of the recombinant kinase was confirmed using model substrates (**f**, lower panels). KD, kinase-dead; RO, RO-3306; BI, BI-2536; ZM, ZM-447439; MBP, myelin basic protein; H1, histone H1. **h**, Local alignment of MCL1 and BAX. Vertical lines and colons mark identical and chemically similar residues, respectively; dashes represent gaps. **i**, Immunoblots of interphase or mitotic Hek293T cells expressing BAX with (+) or without (Δ) TM and with Glu (E), Ala (A), or Thr (T) at position 146. **j**, Model of MOMP by MCL1-C homo-oligomerization. The indicated conformational changes are inspired by knowledge about BAK/BAX pore formation and their hierarchy was chosen to best explain our data; that is, why the T301A mutation abolished self-interaction of MCL1-C but left BIM binding unaffected (Extended Data Fig. 5d). Minus signs represents phosphorylated Thr301 and plus signs represent a nearby basic residue.

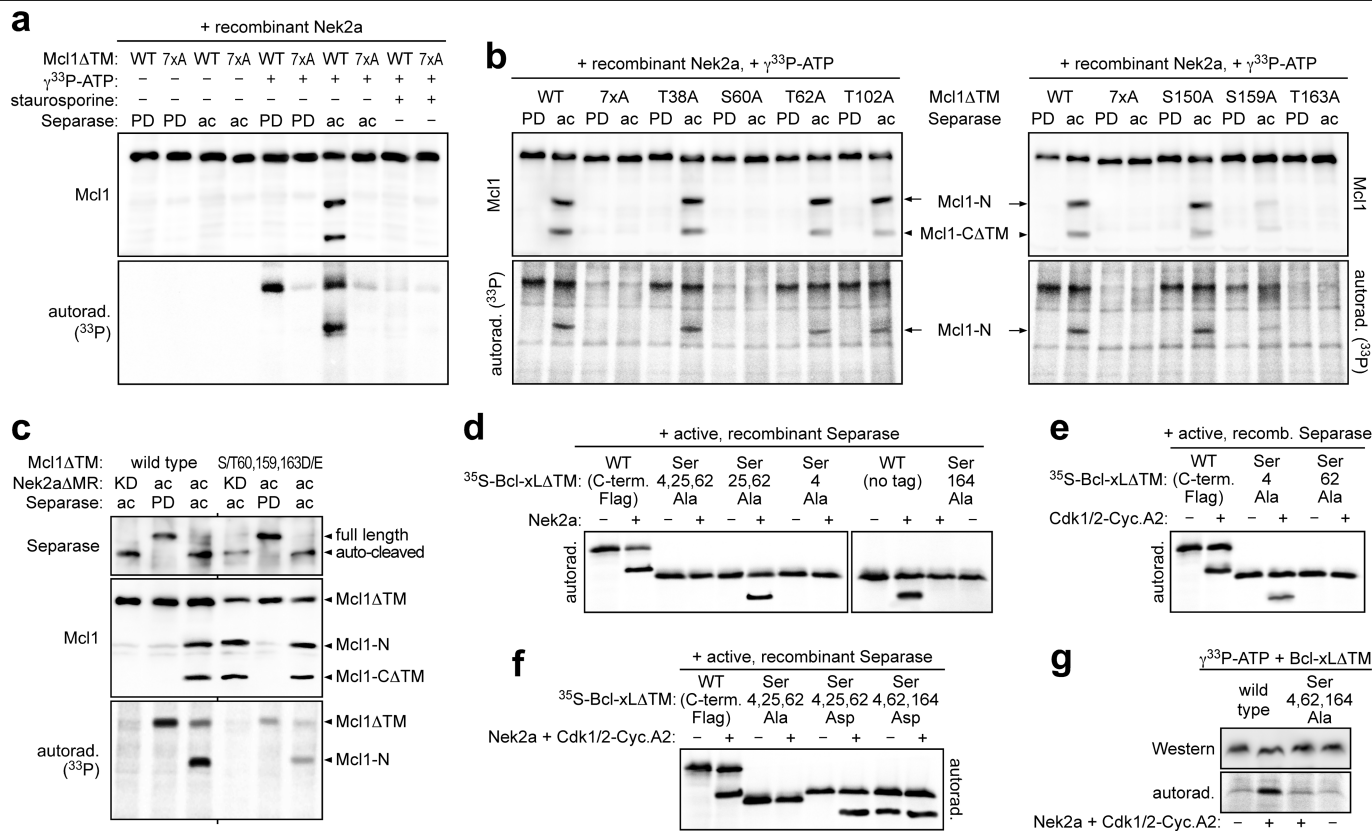


Extended Data Fig. 8 | See next page for caption.

Article

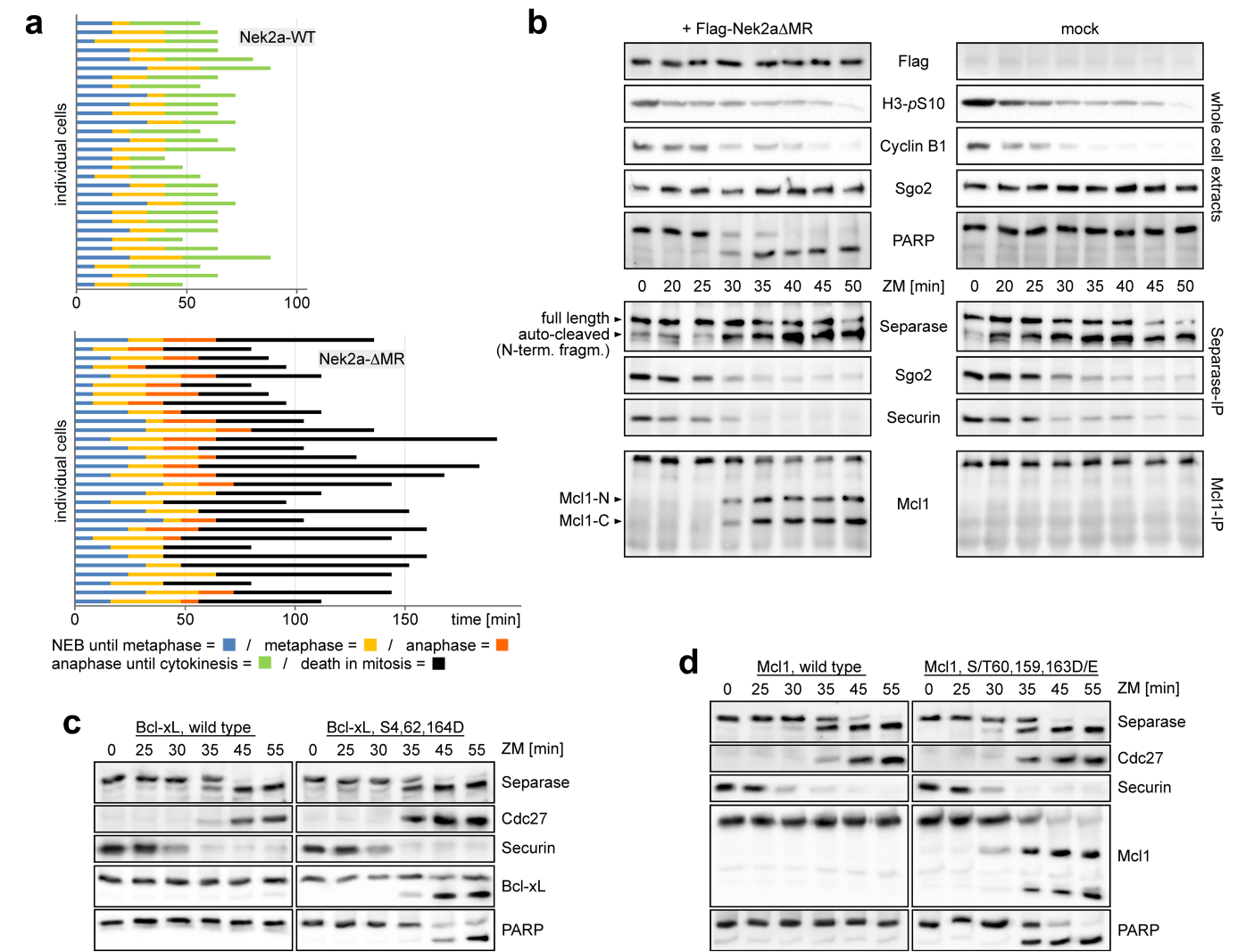
Extended Data Fig. 8 | NEK2A kinase turns MCL1 and BCL-XL into separase substrates. **a**, Autoradiographs and immunoblots of combined kinase (^{33}P -labelling) and cleavage (fragment-generation) assays using bacterially expressed, purified MCL1 ΔTM and kinases, specific inhibitors and active (ac) or inactive (PD) separase, as indicated. **b**, CDK1–cyclin B1 and PLK1 enhance separase-dependent cleavage of NEK2A-phosphorylated MCL1. Immunoblots of cleavage assays using bacterially expressed, purified MCL1 ΔTM and the indicated combination of kinases and separase variants. **c**, When immunoprecipitated from NEK2A- ΔMR -expressing, SAC-arrested cells, endogenous MCL1 is efficiently cleaved by separase in vitro. Left, experimental setup; right, immunoblots of cleavage assay combining active or inactive separase with MCL1 immunoprecipitation from prometaphase Hek293T cells expressing MCL1 ΔTM –Flag and, where indicated, Flag–NEK2A- ΔMR or N-terminally truncated ($\Delta 86$) Flag–cyclin A2. **d**, Cleavage of mouse (*M.m*) MCL1 by separase also requires NEK2A-dependent phosphorylation. In vitro-translated ^{35}S -MYC6-*M.m*.MCL1 ΔTM was incubated with separase,

caspase 3 and NEK2A/ATP as indicated. Reactions were resolved by SDS–PAGE and analysed by autoradiography and immunoblotting. KD, kinase-dead (K37M). **e**, NEK2A and (to a lesser extent) CDK1/2–cyclin A2 sensitize BCL-XL to separase. Autoradiography of cleavage assay combining in vitro-translated ^{35}S -RAD21 (positive control) or ^{35}S -BCL-XL ΔTM with kinases/ATP and separase variants as indicated. **f**, Autoradiography of kinase assays (^{33}P -labelling) using model substrates and the kinases from **d**. **g–i**, The NEK2A-related NEK2B does not support separase-dependent MCL1 cleavage. **g**, Schematics and C-terminal sequences (dashed box) of NEK2A and NEK2B, which arise by alternative splicing of the same gene⁵⁴. NEK2A-specific, C-terminal degrons (KEN box and MR-tail) are underlined. **h**, Both NEK2A and NEK2B can phosphorylate the model substrate MBP. **i**, NEK2B cannot phosphorylate MCL1. Left, experimental setup; top right, kinase assay; bottom right, cleavage assay combining MCL1 ΔTM with NEK2A, NEK2B, staurosporine (kinase inhibitor) and active separase as indicated.



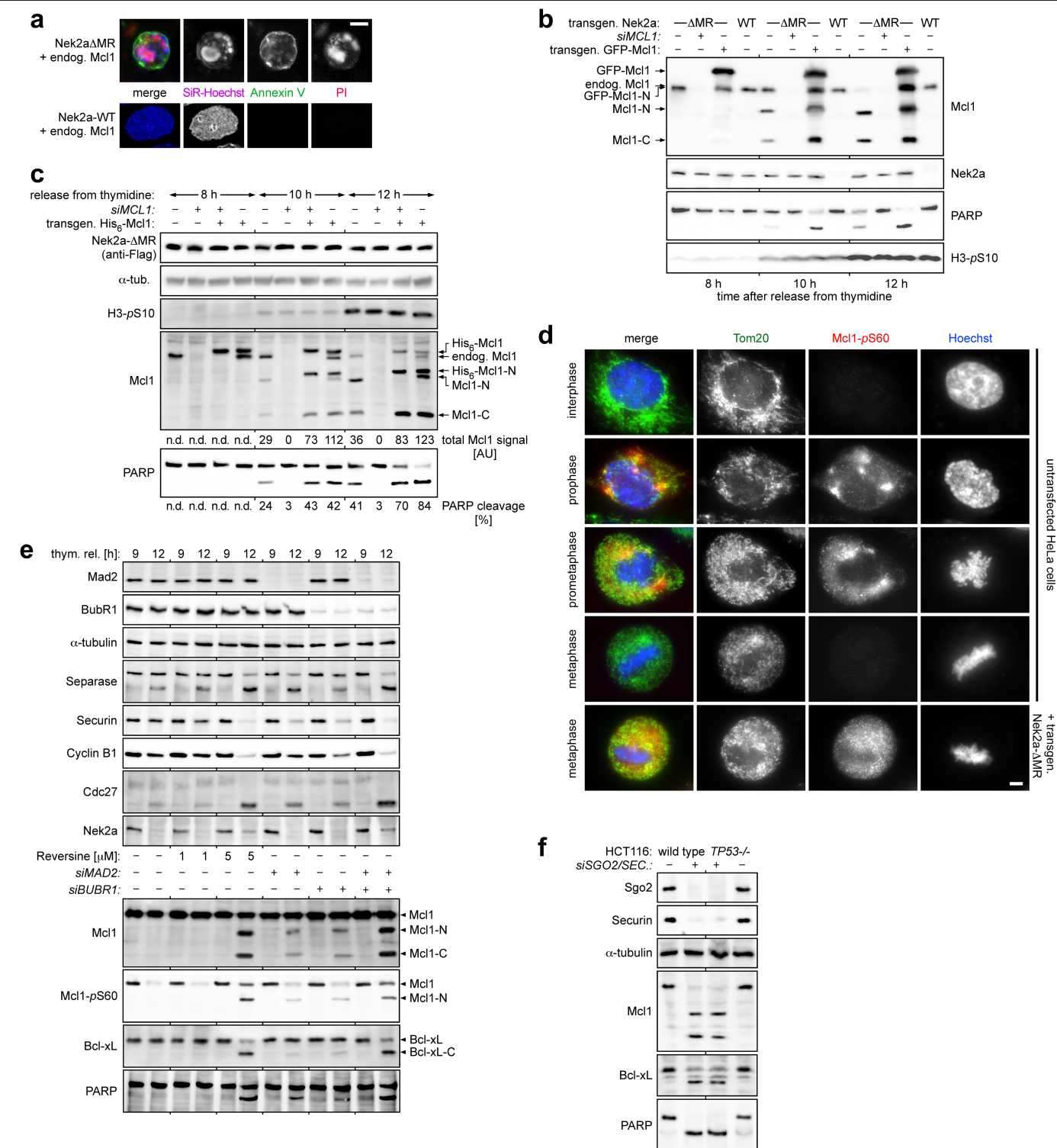
Extended Data Fig. 9 | Mapping cleavage-relevant phosphorylation sites within MCL1 and BCL-XL. a, b, Cleavage of MCL1 by separase essentially requires NEK2A-dependent phosphorylation of Ser60 and Thr163. Autoradiographs and immunoblots of combined kinase (³³P-labelling) and cleavage (fragment-generation) assays. Prior to analysis, in vitro-translated, wild-type MCL1ΔTM and variants thereof were incubated with active NEK2A, γ³³P-ATP, staurosporine, and active (ac) or inactive (PD) separase as indicated. In vivo phosphorylation of Ser159 and Thr163 has previously been reported^{55,56}. In vivo phosphorylation of Ser60 was detected by a phosphorylation-specific antibody (Fig. 4c, Extended Data Fig. 11d, e). **c,** Separase-dependent cleavage of MCL1(S/T60,159,163D/E) is independent of NEK2A. Immunoblots and

autoradiography of combined kinase (³³P) and cleavage assay. **d-f,** In vitro-translated, ³⁵S-labelled wild-type BCL-XLΔTM and variants thereof were incubated with the indicated kinases (+) or reference buffers (-) and active separase before SDS-PAGE and autoradiography. **d,** NEK2A-stimulated cleavage of BCL-XL by separase essentially requires Ser4 and Ser164. **e,** CDK1/2-cyclin A2-stimulated cleavage of BCL-XL by separase essentially requires Ser62. In vivo phosphorylation of Ser62 has been previously reported⁴⁸. **f,** Separase-dependent cleavage of BCL-XL(S4,62,164D) occurs independently of NEK2A and CDK1/2-cyclin A2. **g,** Autoradiograph and immunoblot of kinase assay using in vitro-translated, wild-type BCL-XLΔTM or its S4,62,164A variant and γ³³P-ATP in combination with the indicated kinases (+) or reference buffers (-).



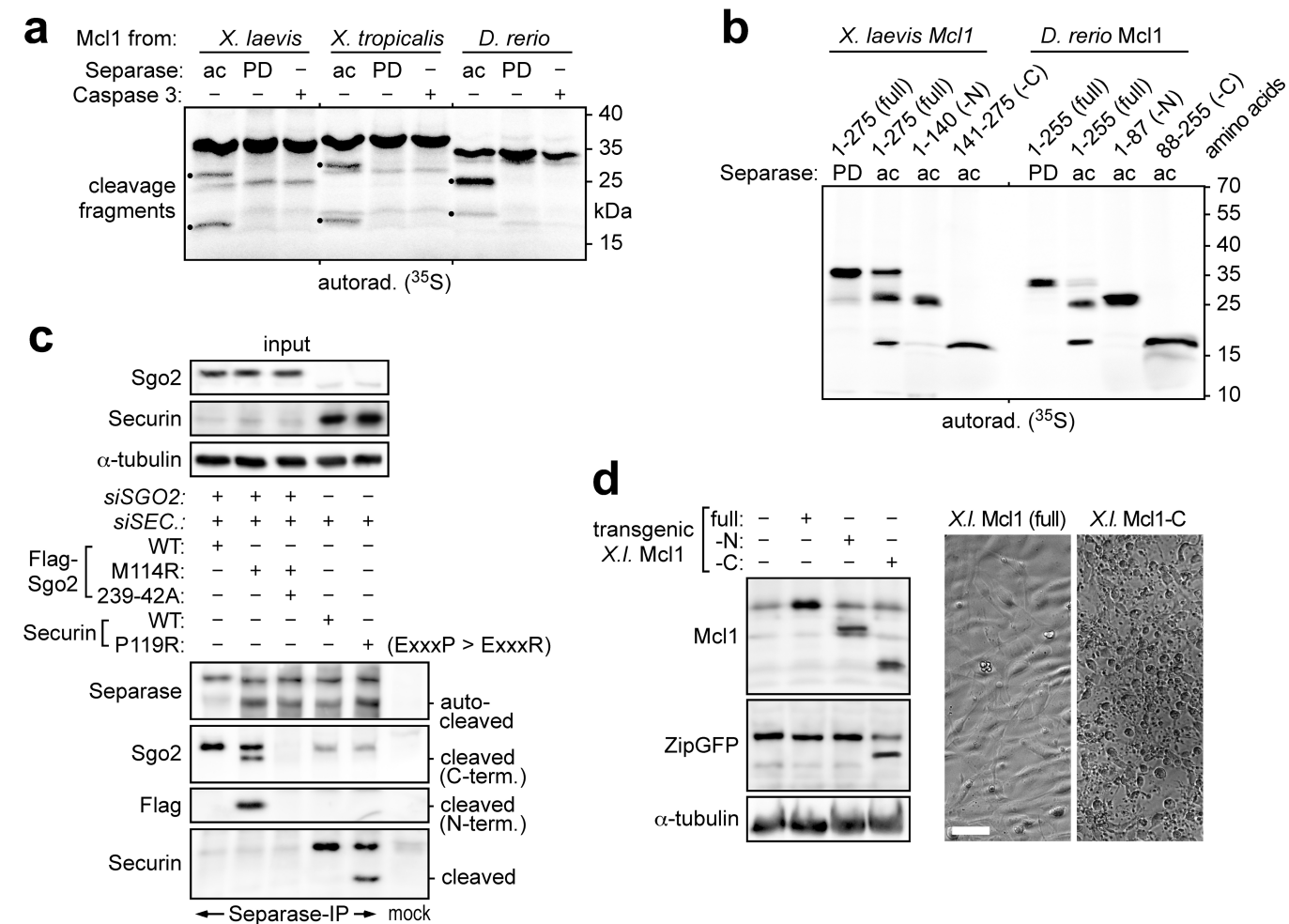
Extended Data Fig. 10 | Stabilized NEK2A and constitutively cleavable MCL1- and BCL-XL variants result in DiMu upon activation of separase in anaphase. a, Cell fate profiles of HeLa-K cells expressing NEK2A-WT or NEK2A-ΔMR and cultured in the presence of SiR-Hoechst and a fluorogenic caspase 3/7 reporter. **b**, Immunoblots of time-resolved separase and MCL1

immunoprecipitations from NEK2A-ΔMR-expressing or control HeLa-K cells released from taxol arrest by addition of ZM-447439 (ZM) at $t = 0$ min. **c, d**, HeLa-K cells expressing the indicated variants of MCL1 and BCL-XL were analysed as in **a**. (Dephosphorylated) CDC27 served as a marker for late mitosis.



Extended Data Fig. 11 | NEK2A stabilization preferentially kills MCL1-overexpressing cells. a, b, Representative images and immunoblots of quantitative analysis shown in Fig. 4b. Scale bar, 5 μm. **c**, Immunoblots of NEK2A-ΔMR-expressing Hek293T cells transfected with siMCL1 and expression plasmids for His₆-MCL1 as indicated. MCL1 and PARP cleavage were quantified densitometrically. n.d., not determined. **d**, MCL1-Ser60 is phosphorylated in early mitosis only. Untransfected or NEK2A-ΔMR-expressing HeLa-K cells were

released from thymidine arrest for 8 h and then analysed by (immuno) fluorescence microscopy using Hoechst and the indicated antibodies. Scale bar, 5 μm. **e**, Chemical abrogation of the SAC triggers DiM. Immunoblots of reversine- or siRNA-treated HeLa-K cells synchronized as in Fig. 4c. Dephosphorylation of CDC27 into a sharp, fast-migrating band serves as a marker of late mitosis. **f**, TP53^{-/-} cells and parental HCT116 cells were depleted of SGO2 and securin, taxol-arrested and analysed by immunoblotting.



Extended Data Fig. 12 | MCL1 cleavage by separase and the pro-apoptotic effect of MCL1-C is conserved in non-mammalian vertebrates.
a, ³⁵S-labelled, NEK2A/ATP-treated full-length Mcl1 from *X. laevis*, *X. tropicalis* and *D. rerio* were incubated with separase variants and caspase 3 as indicated, and analysed by autoradiography. **b**, **c**, Separase can cleave after ExxxR motifs. **b**, *X. laevis* and *D. rerio* Mcl1 are cleaved by separase after 136-ExxxR-140 and 84-ExxR-87, respectively. The indicated full-length Mcl1 or fragments thereof

were analysed as in **a**, **c**. Changing the pseudo-substrate sequence of securin to ExxxR turns it into a separase substrate. Endogenous human SGO2 and securin were depleted by RNAi and replaced by the indicated variants⁹. These were then assessed for cleavage by immunoprecipitation and western blotting. **d**, *Xenopus* S3 cells transfected to express ZipGFP and the indicated forms of *X. laevis* Mcl1 were analysed by immunoblotting (left) and video microscopy (representative phase contrast images on right). Scale bar, 50 μm.

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- ☒ ☒ The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- ☒ ☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- ☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- ☒ ☐ A description of all covariates tested
- ☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- ☐ ☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- ☒ ☐ For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- ☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- ☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- ☒ ☐ Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

No commercial open source software was used to collect data.

Data analysis

Because no complex statistical analysis is performed standard analysis programmes (excel 2016, sigma Plot 9.0) were used. Densitometric quantification of western blot signals was done using Multi Gauge (Fujifilm). For FACS data analysis CXP 2.2 software (Beckman Coulter) was used. To handle and edit live cell videos and images LasAF version 2.7.0.9329 (Leica) was utilized. For 2D-SIM image reconstruction ImageJ 1.52i was used.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The authors declare that all data supporting the findings of this study are available within the paper and its supplementary information files. If there is reasonable request data can also be provided from the corresponding author.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We did not perform statistical analysis to determine a specific sample size. We used standard cell culture based experiments and minimally reproduced the results in three independent trials using cells from individual cryo-stocks.
Data exclusions	No data were excluded.
Replication	The values obtained from distinct experimental trials were reproducible. The data are presented as means together with the corresponding individual data points from each repetition to indicate biological variation. Experiments analysed by immunoblotting were repeated 2-4 times with similar results (2-4 biological replicates).
Randomization	n/a
Blinding	For quantitative analyses of chromosome spreads and IFM specimen the investigators were blinded to sample allocation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

Methods

n/a	Involved in the study
<input type="checkbox"/>	<input checked="" type="checkbox"/> Antibodies
<input type="checkbox"/>	<input checked="" type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input checked="" type="checkbox"/>	<input type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Antibodies

Antibodies used

Rabbit anti-separase (1.5 µg/ml) see Ref. in Material and Methods, mouse anti-Flag M2 (1:2,000; Product No. F1804; Sigma-Aldrich), rabbit anti-Sgo2 (1:1,000; Product No. A301-262A; Lot. No. A301-262A-1; Bethyl), rabbit or guinea pig anti-Sgo2 (1.5 µg/ml; raised by Charles river laboratories against the peptide: DVPPRESHSQSSK), rabbit anti-Sgo1 (1:500, Abcam ab21633), mouse anti-securin (1:1,000; clone DCS-280; Code No. K009-3; MBL), rabbit anti-phosphoSer10-histone H3 (1:1,000; Product No. 06-570; Lot No. 2370127; Millipore), mouse anti-cyclin B1 (1:1,000; Product No. 05-373; Lot. No. 2199734; Millipore), rabbit anti-Pin1 (1:1,000) see Ref. in Material and Methods, rabbit anti-caspase 3 Asp175 (1:1,000; clone 5A1E; Product No. 9664; Cell Signaling), rabbit anti-Bcl-xL (1:1,000; Product No. 2762; Cell Signaling), mouse anti-Mcl1 (1:800; clone W16014A; Product No. 695702; BioLegend), guinea-pig anti-Mcl1 (1 µg/ml; raised by Charles River Laboratories against bacterially expressed human Mcl1ΔTM), guinea pig anti-phosphoSer10-Mcl1 ('Mcl1-pS60'; 0.5 µg/ml; for IFM 1 µg/ml; anti-CVIGGpSAGA liberated from reactivity towards CVIGGSAGA), rabbit anti-Sororin (1 µg/ml; raised by Charles River Laboratories against bacterially expressed full-length human sororin), rabbit anti-PARP (1:800; clone 46D11; Product No.9532; Cell Signaling), mouse anti-PARP (1:1,000; clone C-2-10; Product No. AM30; Calbiochem), mouse anti-Tom20 (1:500; Santa-Cruz Biotechnology, F10), mouse anti-cytochrome c (1:1,000; BD Pharmingen, 7H8.2C12), mouse anti-BubR1 (1:1,000; BD Transduction Laboratories, clone 9), rabbit anti-Bax (1:1,000; Abcam, ab32503), rabbit anti-Bak (1:1,000; Abcam, ab32371), rabbit anti-Phosphothreonine (1 µg/ml; Product No. 71-8200; Invitrogen), rabbit anti-Mad2 (1:1,000; Bethyl, A300-300A), anti-MBP monoclonal (1:1,000; NEB Biolabs; HRP-conjugated; Product No. E8032), mouse anti-Nek2 (1:600; clone 20/Nek2 Ruo; Product No. 610593; BD Transduction Laboratories), goat anti-Cdc27 (1:1,000) see Ref. in Material and Methods, mouse anti-RGS-His (1:1,000; Product No. 34610; Qiagen), rabbit anti-phosphoSer139-histone H2A.X (γH2AX; 1:5,000; clone EP854(2)Y; Product No. MABE205; Lot No. 2034733; Millipore), mouse anti-topoisomerase IIα (1:1,000; clone 1C5; Product No. ADI-KAM-CC210-E; Enzo Life Sciences), mouse anti-cyclin A2 (1:200; clone 46B11; Product No. sc-53234; Santa Cruz Biotechnology), rat anti-HA (1:2,000; clone 3F10; Product No. 11867423001; Roche), mouse anti-GFP (hybridoma supernatant 1:2,000; gift from D. van Essen and S. Sacconi), and mouse anti-α-tubulin (hybridoma supernatant 1:200; DSHB; clone 12G10), rabbit anti-murine Pttg1 and anti-murine SgoL2 sera (both 1:1,000; gift from Alberto M. Pendas).

Validation

Validation procedures used for commercial antibodies are described by the manufacturer. Self-made polyclonal rabbit or guinea pig antibodies are validated using multiple methods (e. g. ICC/IF, WB or IP). To address antibody specificity we used the corresponding antigen (e.g. recombinantly expressed in E. coli, IVTT or cell lysate) as positive and siRNA depleted lysate or fixed cells as negative control.

Eukaryotic cell lines

Policy information about [cell lines](#)

Cell line source(s)

Hek293T (Gift from Marc W. Kirschner); HeLa-K (Gift from Thomas U. Mayer); RPE-1 hTERT (ATCC, CRL-4000); HCT116 parental and Bak/Bax DKO (Gift from Richard Joule); SW480, T47D, A427 and A549 (Gift of Michael Orth from); NIH3T3 purchased from ATCC (CRL-1658); TP53 KO (Gift from Bert Vogelstein); Xenopus laevis S3 (Gift from Guowei Fang)

Authentication

Cell were authenticated via visual inspection of typical morphology, by immunoblotting analyses and cell synchronization behavior, and through selective resistance to antibiotic treatments.

Mycoplasma contamination

Cell lines were not tested for mycoplasma contamination but microscopic inspections of their fluorescently labeled DNA contents were inconspicuous.

Commonly misidentified lines
(See [ICLAC](#) register)

No commonly misidentified cell lines were used.

Author Correction: Self-verifying variational quantum simulation of lattice models

<https://doi.org/10.1038/s41586-020-2203-2>

Correction to: *Nature* <https://doi.org/10.1038/s41586-019-1177-4>

Published online 15 May 2019



Check for updates

C. Kokail, C. Maier, R. van Bijnen, T. Brydges, M. K. Joshi, P. Jurcevic, C. A. Muschik, P. Silvi, R. Blatt, C. F. Roos & P. Zoller

In the Acknowledgements of this Article, the text “and the Quantum Flagship PASQUANS” should be replaced with the sentence “This project (or publication) has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement no. 817482 (PASQuanS).” The original Article has been corrected online.

Publisher Correction: Z-nucleic-acid sensing triggers ZBP1-dependent necroptosis and inflammation

<https://doi.org/10.1038/s41586-020-2207-y>

Correction to: *Nature* <https://doi.org/10.1038/s41586-020-2129-8>

Published online 25 March 2020



Check for updates

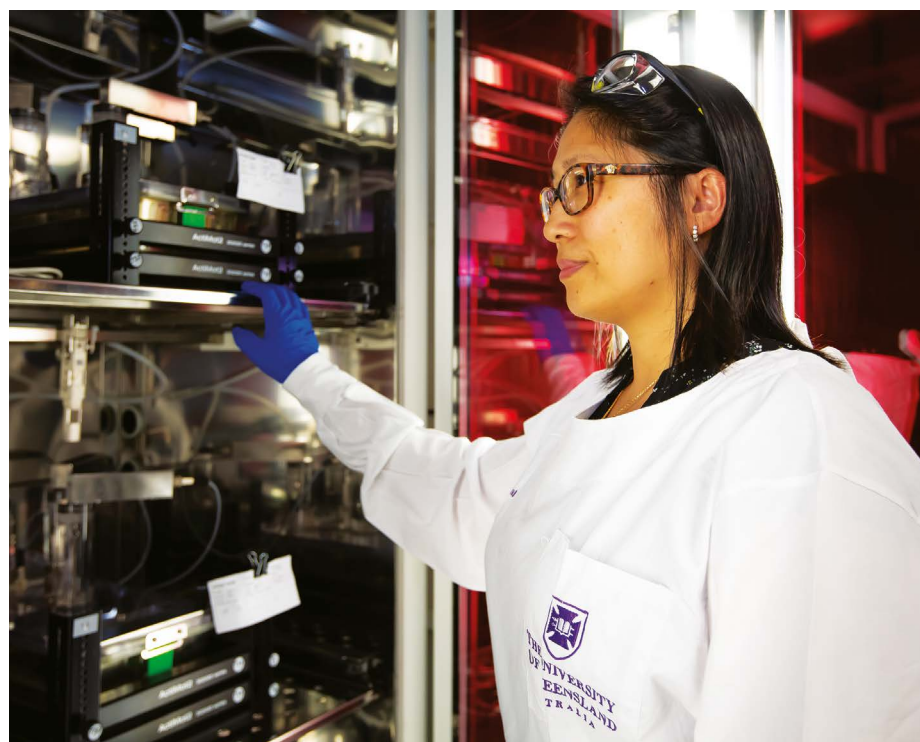
Huipeng Jiao, Laurens Wachsmuth, Snehlata Kumari,
Robin Schwarzer, Juan Lin, Remzi Onur Eren, Amanda Fisher,
Rebecca Lane, George R. Young, George Kassiotis, William J. Kaiser
& Manolis Pasparakis

In Fig. 1b of this Article, owing to errors in the production process, there is an incorrect P value in the left-hand graph of this panel. The value ' $P < 10^{-4}$ ', corresponding to the comparison between control and RIPK1^{E-KO}, should be ' $P = 2 \times 10^{-4}$ '. In addition, the label 'RIPK1^{E-KO}' was duplicated in Fig. 1c. These errors have been corrected online.

Work

Your
story

Send your careers story
to: naturecareerseditor@nature.com



Shyuan Ngo and her lab members work in rotating groups to maintain distance.

KEEPING ON WITH ESSENTIAL SCIENCE

For scientists working on urgent research, staying home is not an option. **By Virginia Gewin**

Around the world, universities have closed because of COVID-19 – forcing an increasing number of researchers to work and teach from home.

Some scientists can't simply stop going to their laboratories – especially not those who are overseeing clinical trials that could offer life-saving vaccines and therapies, particularly against the new coronavirus. And some research activities unrelated to vaccine production must continue, even in the face of an institutional shutdown.

"Animals need to be looked after, and breeding lines must be kept going. Many of these are unique and can't be regenerated," says Mike Turner, director of science at Wellcome, a research-funding charity in London.

If you have to go to your lab, says Turner, comply with all of your institution's safety

regulations, not just those for preventing COVID-19. For example, stick to the 'buddy system' – by working in groups of at least two people – when necessary for your safety. Many current guidelines advise a person not to come within 2 metres of anyone for longer than 15 minutes.

Turner also suggests creating schedules to make sure the essential tasks of each lab at an institution get done if COVID-19 strikes among the members in any one of the labs.

"If someone gets ill, they know who to tell, and that person knows it is their role to continue the task," he says. "If your work is contributing to [research against] COVID-19, we applaud you, and please carry on. If it's not, ask, 'Is it absolutely essential?'"

Here, four scientists offer advice on the precautionary measures necessary to

continue essential research in the face of the pandemic.

JOHN MORRISON REDESIGN STUDIES TO BE LESS LABOUR-INTENSIVE

The status of the lab changes every day. In California, we have a statewide order to 'shelter in place', or to stay home except to do essential tasks, such as seeking health care or buying groceries. My colleagues and I are exempt from that order because we are beginning aggressive COVID-19 research in collaboration with the Center for Immunology and Infectious Diseases at the University of California, Davis, and we care for animals.

We are exercising an extreme form of social distancing while at work. It's complicated for our group. We have to worry about the health of our people – our primary concern – as well as the health of the primate colony we use for research. We have to assume that there could be human-monkey transmission and vice versa. And we have to maintain the colony.

We are not starting any new protocols or studies beyond COVID-19 research, and are doing what we can to keep existing research protocols going. We decided to give preference to longitudinal studies, to make sure data collected previously remain meaningful. We are asking people to redesign their studies to require 50% less labour while salvaging the study and making it rigorous and reproducible. It can be terribly difficult to decide which studies must have data collection stopped. It's important to keep the measures going that yield the biggest bang for the buck, or the most information for the least effort and lowest cost. For example, can you collect samples less often without compromising the validity of the study? In some instances, the smartest thing to do might be to accelerate the study's completion to get it out of the way.

We have 300 people here, and are reducing our on-site staff by at least 50%. But we are the only primate-research centre growing the new coronavirus to help develop clinical tests. People are working from home, staggering their shifts and spreading out across labs. What gets hard is launching new COVID-19 research with our reduced capacity. And it became very clear that most of the national primate-research centres are initiating COVID-19 research as a high priority. We're doing that because the monkey model for the disease will be extraordinarily powerful, and we need it yesterday.

Work/Careers

Protecting our workforce is our highest priority, followed closely by helping to fight this pandemic while keep our monkeys safe.

John Morrison is director, California National Primate Research Center at the University of California, Davis.

VIJIVIJAYAN COLLECTIVELY SACRIFICE TO KEEP WORK GOING

Our university has not yet considered shutting down, but we encourage those who can to work from home. For those continuing wet-bench lab work, we follow the clear advice that we receive from Singapore's Ministry of Health and our university administrators on social-distancing guidelines and on specific directives, such as a 14-day self-quarantine for people arriving from certain countries. Fortunately, the university already had a robust biorisk-management system in place, which made it easier to beef up our response to the COVID-19 outbreak. We are now taking precautions to make sure work goes on, because not everyone can work remotely.

During the first couple of days of the outbreak, our infectious-disease researchers were working long hours. We reduced those,

"Our work isn't about academic credit right now — it's about what we can do for our country and the world."

because it is not safe to work while fatigued. We now make sure that working hours are within acceptable limits so that people don't get tired — and have time to take breaks.

There are about 400 researchers at our medical school across roughly 50 wet-bench labs. Roughly 20% of them work on infectious disease, and the others work on cancer, metabolic disorders and neurobehaviour. We have thinned our workplace staff by reducing the number of people working in labs by about 40–50%. Labs have broken into teams, and have adopted a variety of strategies to distance workers in both time and space.

Our biosafety-level 3 (BSL-3; level 4 is the most strict) lab is small, so researchers working in that space have split into early-morning and late-afternoon shifts. In other labs, teams either switch days of each week — working Monday to Wednesday or Thursday and Friday — or split into different floors of the building, depending on what works best for their research. That way, if one group is quarantined, the other group can take over. The teams each wear differently coloured stickers all the

time to avoid members from other teams.

In addition, given the increased demand for the small BSL-3 space, researchers can make non-infectious coronavirus by extracting its RNA so they can work with it, when feasible, in the less-stringent BSL-2 lab environments.

Continuing lab experiments is possible, but not easy. It's a collective sacrifice to make big changes in the way you work and live, but these are absolutely essential to combat this outbreak. Our biggest asset is our people at work.

Viji Vijayan is associate dean of safety and emergency management, Duke–National University of Singapore Medical School.

SHYUAN NGO LET ETHICS GUIDE WHICH RESEARCH TO CONTINUE

My lab of ten includes postdoctoral researchers, PhD students and research assistants. Normally, we do natural-history studies with people with motor-neuron disease (MND) in a clinical setting, and take everything we learn back to the preclinical laboratory, using mouse models of MND and stem cells from people with the condition. Right now, the University of Queensland remains open, although it recently announced a move to teaching online and advised against non-essential travel.

Research-group leaders were asked to develop contingency plans to reduce the number of people in their labs. My lab is one of the few on campus that has adopted those plans. We shut down clinical research because our participants are at risk of respiratory distress from COVID-19. We also stopped projects that were initiated in January, choosing instead to focus our efforts on projects that were closer to completion. It would be unethical to cull animals that we have already been working with for almost a year. Similarly, we're invested in making sure we get all the data necessary from stem cells grown from people with MND, because they have donated their time and invested in the project. And this research was funded by publicly donated money — another reason not to shut it down. For safety, we divided the lab into two groups, each composed of four individuals. Within each team, we rotate the schedule, leaving a 30-minute gap between shifts, so that people never cross paths in the lab. We do this throughout the open space we share with three other labs. The stem-cell work happens in the early morning and late afternoon, so that the animal team can continue collecting samples at the same time it has been doing every day, for continuity.

Consult with your whole team on how to prioritize projects and schedules and involve everyone in decision-making, to avoid a sense of uncertainty and disgruntlement. We don't

need that right now. We need to all work together.

Shyuan Ngo is an MND researcher at the University of Queensland in Brisbane, Australia

MARK DENISON COMMUNICATE CONSTANTLY

I started working on coronaviruses in 1984, and have worked through severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS). I am sobered and subdued by the reality of the COVID-19 pandemic. Our team's priority is work on countermeasures.

We are conducting *in vitro* studies of potential antiviral drugs including remdesivir, alone or in combination with other compounds, while continuing our long-standing collaboration with Ralph Baric's lab at the University of North Carolina in Chapel Hill. Theat group is developing animal models to test these potential COVID-19 therapies. We expect to continue participating in collaborations to test vaccines. My goals are making sure nothing hampers the speed of vaccine trials, and identifying any drug combinations that could have a high impact in mitigating the disease.

Our work isn't about academic credit right now — it's about what we can do for our country and the world. We want to contribute swiftly, so I'm working out how to do that with what we have. Already, the university has recognized our essential work and commitment to safety, and has allowed postdocs and graduate students to continue working in our lab.

I have a 12-member team; 6, including me, are trained to work in a BSL-3 lab. We work long days, seven days a week. I try not to do that, but people in the emergency rooms in Italy are working more than that — and they don't have a choice.

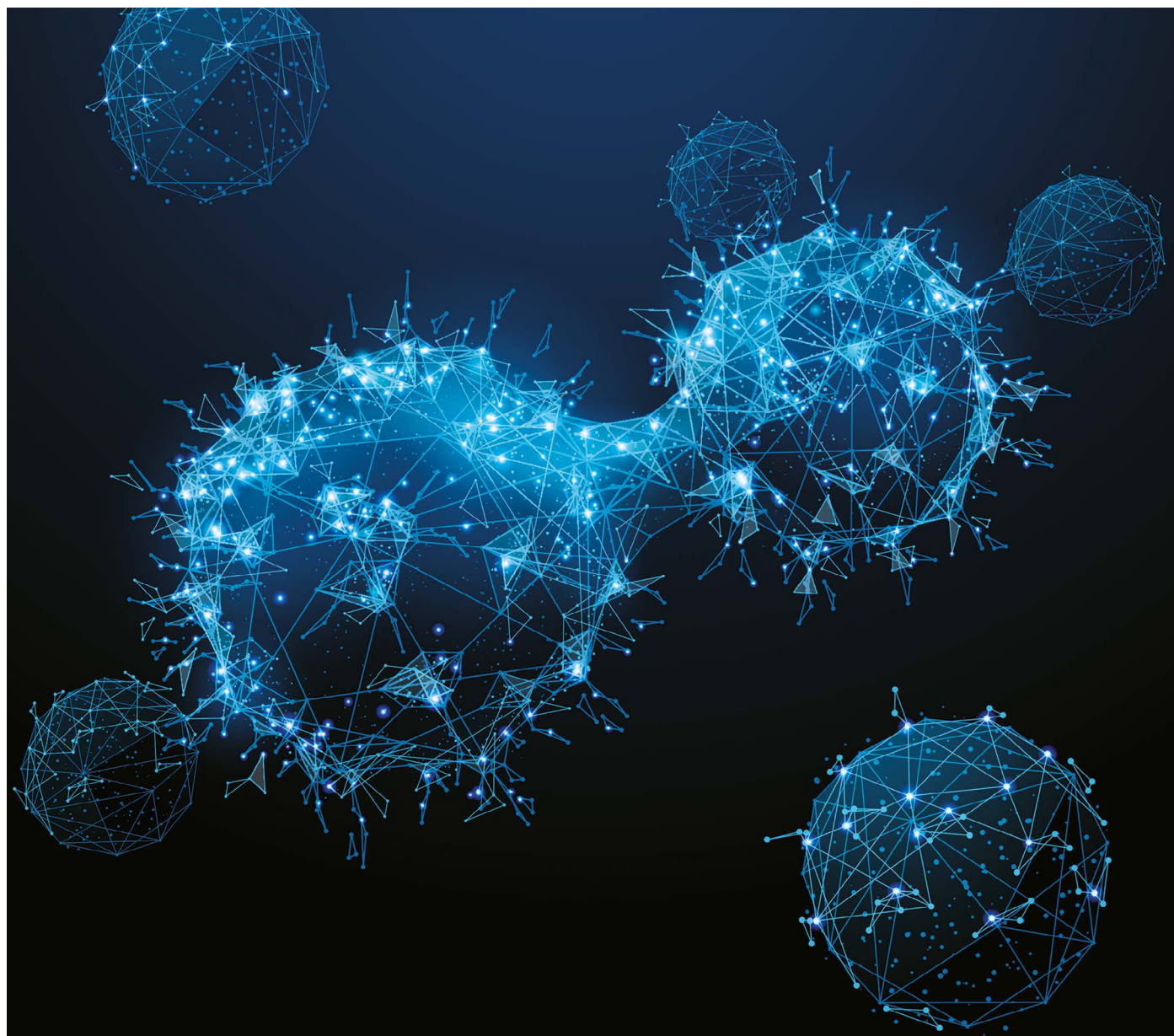
At the same time, we can't get so tired that we make mistakes. We can't work all the time. We take steps to protect ourselves and each other. Everyone absolutely has to report any symptoms of any kind — from an earache to a runny nose. We are in constant contact.

To keep everyone more than 2 metres apart, we have got people working in other labs that shut down. We conduct our staff meetings by talking across the halls from individual labs, or through the videoconferencing platform Zoom. If one of us gets COVID-19, the whole research programme could shut down.

Mark Denison is director of paediatric infectious diseases, Vanderbilt University Medical Center, Nashville, Tennessee.

Interviews by Virginia Gewin

These interviews have been edited for length and clarity.



SHUTTERSTOCK

DEEP LEARNING TAKES ON TUMOURS

Artificial-intelligence methods are moving into cancer research. **By Esther Landhuis**

As cancer cells spread in a culture dish, Guillaume Jacquemet is watching. The cell movements hold clues to how drugs or gene variants might affect the spread of tumours in the body, and he is tracking the nucleus of each cell in frame after frame of time-lapse microscopy films. But because he has generated about 500 films, each with 120 frames and 200–300 cells per frame, that analysis

is challenging to say the least. “If I had to do the tracking manually, it would be impossible,” says Jacquemet, a cell biologist at Åbo Akademi University in Turku, Finland.

So he has trained a machine to spot the nuclei instead. Jacquemet uses methods available on a platform called ZeroCostDL4Mic, part of a growing collection of resources aimed at making artificial intelligence (AI) technology accessible to bench scientists who have

minimal coding experience¹.

AI technologies encompass several methods. One, called machine learning, uses data that have been manually preprocessed and makes predictions according to what the AI learns. Deep learning, by contrast, can identify complex patterns in raw data. It is used in self-driving cars, speech-recognition software, game-playing computers – and to spot cell nuclei in massive microscopy data sets.

Deep learning has its origins in the 1940s, when scientists built a computer model that was organized in interconnected layers, like neurons in the human brain. Decades later, researchers taught these ‘neural networks’ to recognize shapes, words and numbers. But it wasn’t until about five years ago that deep learning began to gain traction in biology and medicine.

A major driving force has been the explosive growth of life-sciences data. With modern gene-sequencing technologies, a single experiment can produce gigabytes of information. The Cancer Genome Atlas, launched in 2006, has collected information on tens of thousands of samples spanning 33 cancer types; the data exceed 2.5 petabytes (1 petabyte is 1 million gigabytes). And advances in tissue labelling and automated microscopy are generating complex imaging data faster than researchers can possibly mine them. “There’s definitely a revolution going on,” says Emma Lundberg, a bioengineer at the KTH Royal Institute of Technology in Stockholm.

Boosting image-based profiling

Cancer biologist Neil Carragher caught his first glimpse of this revolution in 2004. He was leading a team at AstraZeneca in Loughborough, UK, that explores new technologies for the life sciences, when he came across a study that made the company rethink its drug-screening efforts. He and his team had been using cell-based screens to look for promising drug candidates, but hits were hard to

come by. The study was suggesting that AI and analytics could help them to improve their screening processes². “We thought this could be a solution to the productivity crisis,” Carragher says.

But AI technologies can be difficult for biologists to master. Jacquemet says he once spent more than a week trying to install the correct software libraries to run a deep-learning model. Then, he says, “you need to learn to code in Python” to use it.

Carragher’s AstraZeneca team worked with computational biologist Anne Carpenter and

“If I had to do the tracking manually, it would be impossible.”

her colleagues at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts, to scale up the image-profiling method used in the 2004 paper and to investigate the effects of multiple drugs on human breast-cancer cells³. Carpenter went on to develop the technique into a procedure called Cell Painting, which stains cells with a panel of fluorescent dyes and then uses the open-source software CellProfiler to generate profiles of the cells.

Still, these analyses can be labour-intensive, says Carragher, who now heads cancer-drug discovery at the University of Edinburgh, UK. Even with open-source tools that avoided the need to code the machine

learning from scratch – and a computing cluster with thousands of processors and terabytes of memory – it could take a month or so to work out which cellular features they should tell the image-analysis software to look at, Carragher says. And after optimizing the parameters for each cell line, his team had to tinker further to get it to work across all cells.

Last year, he and his team explored how deep learning could improve this process. The impetus was a 2017 analysis⁴ posted on the bioRxiv preprint server by researchers at Google’s headquarters in Mountain View, California. The researchers had downloaded Carragher’s breast-cancer data set from the Broad Bioimage Benchmark Collection and used it to train a deep neural network that previously had seen only general images, such as cars and animals. By scanning for patterns in the breast-cancer data, the model learnt to discern cellular changes that are meaningful for drug discovery. Because the software wasn’t told what to look for, it found features that researchers hadn’t even considered.

Building on that effort, Carragher and his colleagues screened 14,000 compounds across 8 forms of breast cancer⁵. “We did identify some interesting hits,” he says – including a compound that was already known to modulate receptors for serotonin, which is important in mammary-gland development, as they reported earlier this year⁶.

At the Broad Institute, a team led by computational biologist Juan Caicedo is applying image-based profiling to screen for genetic mutations. He and his team overexpressed various gene variants in lung-cancer cells, stained them with the Cell Painting protocol and looked for differences in the cells that suggest possible pharmaceutical opportunities. They found that machine learning could identify meaningful variants in images about as well as processes that measure gene expression in the cells. The researchers reported their results at the AI Powered Drug Discovery and Manufacturing Conference in February at the Massachusetts Institute of Technology in Cambridge.

As part of the Cancer Cell Map Initiative, which maps molecular networks underlying human cancer, researchers are training a deep-learning model to predict drug responses on the basis of a person’s cancer-genome sequence. Such predictions have life-or-death implications, and accuracy is crucial, says Trey Ideker, a bioengineer at the University of California, San Diego. But some are reluctant to accept results when the mechanisms behind them aren’t clear, and deep neural networks produce answers without revealing their process – a problem known as ‘black-box’ learning. “You want to know why,” says Ideker. “You want to know the mechanism.” Ideker’s team is creating

WANTED: MORE DATA

Deep-learning models can process raw data, but first they must be trained with annotated information.

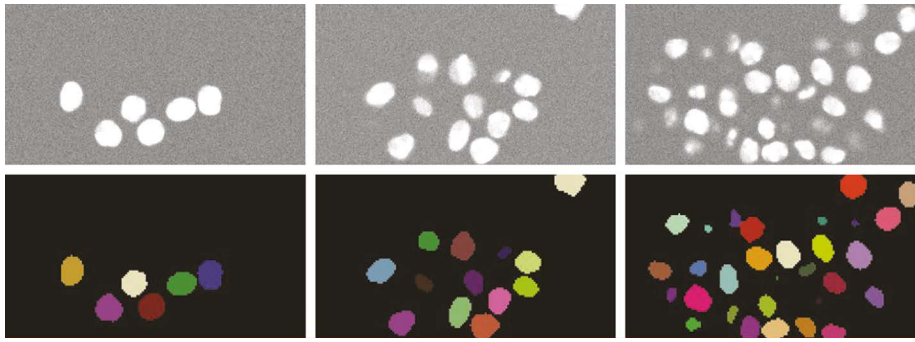
It takes vast amounts of labelled data to train deep-learning models. But that’s not always easy to come by, says Casey Greene, a computational biologist at the University of Pennsylvania in Philadelphia. “Data are cheap, but labelled data are expensive.”

In the genomics realm, sequences are abundant and publicly available. But their associated descriptions, or metadata, are often missing, wrong or unstandardized, says Emily Flynn, a doctoral candidate in biomedical informatics at Stanford University in California. A researcher wanting to train a model to detect non-small-cell lung cancer in samples from patients, for example, might well find data sets variously labelled ‘nsclo’, ‘non small-cell’ or ‘non small cell LC’ – differences that confound analysis tools. Or samples might be labelled ‘disease:

glioblastoma’ and ‘disease: yes’, says biostatistician Colin Dewey at the University of Wisconsin–Madison.

To help organize those data, Dewey created a computational pipeline called MetaSRA, which uses text-mining techniques to standardize and store metadata on public sequences. And Greene and colleagues have built refine.bio, a repository that harmonizes data on expression and RNA sequencing. Working with Stanford bioengineer Russ Altman, Flynn is using machine-learning techniques to infer missing labels from gene-expression data to improve annotations in refine.bio.

In bioimaging, the problem lies more in annotation. To label a set of histopathology slides, for example, “someone has to go in and draw a bounding box around the parts that are cancer”, Greene says. “And that person probably makes a lot of money.” Now developers are training deep-learning algorithms to label nuclei and other structures in cell images, while the Image Data Resource and other online repositories are making it easier for researchers to share and find life-sciences images.



Cell nuclei (top, DNA stain) are automatically detected using the CellProfiler method (bottom).

a ‘visible’ neural network, which links the model’s inner workings more directly to cancer cell biology. As a proof of concept, the team created a model for yeast cells. Called DCell, it can predict the effects of gene mutations on cell growth and the molecular pathways underlying those effects⁷.

The spatial dimension

Lundberg and others in Sweden are using deep learning to tackle another computational challenge: assessing protein localization. The work is part of the Human Protein Atlas, a multi-year, multi-omics effort to map all human proteins. Spatial information reveals where proteins are located in cells, and tend to be under-represented in systems-level studies, Lundberg says. But if researchers knew this information, they could use it to glean insights about the underlying biology, she suggests.

Enter AI. In 2016, Lundberg and her colleagues invited gamers to help computers classify proteins’ whereabouts in cells. The citizen scientists took part in a role-playing game called EVE Online, in which they had to pinpoint fluorescently labelled proteins to win game credits, boosting an AI system already used for this purpose. But even the augmented system trailed human experts in terms of accuracy and speed.

So, in 2018, Lundberg’s team took its images to Kaggle – a platform that challenges machine-learning experts to develop their best models to crack data sets posted by companies and researchers. Over the course of 3 months, 2,172 teams around the world competed to develop a deep-learning model that could look at a cell stained for a protein and several reference markers, and work out the protein’s spatial distribution.

The task was challenging. Half of human proteins are found in multiple places in cells, says Lundberg. And some cellular compartments – the nucleus, for example – are much more common locations than others.

Still, the Kagglers delivered, Lundberg says. Most of the leading strategies came from computational scientists with no biology background – including Bojan Tunguz, a software engineer who created models that predict earthquakes and loan defaults before

earning one of the top spots in the Human Protein Atlas contest. The approach to these problems is similar across vastly different disciplines, Tunguz says.

The best model identified both rare and common locations across a variety of cell lines and, most importantly, captured mixed patterns well, Lundberg says. The algorithm performed almost as accurately as human experts, and with greater speed and reproducibility. Furthermore, it could quantify the spatial information⁸. “When we can quantify it, and not just describe it with a label, we can integrate it with other types of data.” That includes ‘omics’ data, which are already transforming cancer research.

A computational framework known as DeepProg applies deep learning to ‘omics’ data sets, including gene expression and epigenetic data, to predict patient survival, for instance⁹. And DigitalDLSorter predicts outcomes by

“When we can quantify it, and not just describe it with a label, we can integrate it with other types of data.”

inferring types and quantities of immune cells directly from tumour-RNA sequencing data rather than relying on laborious conventional workflows¹⁰.

On the horizon

Many of the tools needed to build deep-learning models are freely available online, including software libraries and coding frameworks such as TensorFlow, Pytorch, Keras and Caffe. Researchers wanting to ask questions and brainstorm solutions to problems that crop up with image-analysis tools can make use of an online resource called the Scientific Community Image Forum (<https://forum.image.sc>). Also becoming available are repositories that allow researchers to find and repurpose deep-learning models for related tasks – a process called transfer learning. One example is Kipoi, which allows researchers to search and explore more than 2,000 ready-to-use models trained for tasks

such as predicting how proteins known as transcription factors will bind to DNA, or where enzymes are likely to splice the genetic code.

Working with other tool developers, Lundberg’s team put together a rudimentary ‘model zoo’ (<https://bioimage.io>) to quickly share its Human Protein Atlas models, and is now creating a more sophisticated repository that will be useful to model producers and non-expert users alike.

A platform called ImJoy will be part of this effort, Lundberg says. Created by Wei Ouyang, a postdoc in her lab, the platform lets researchers test and run AI models through a web browser on their computer, in the cloud or on a phone. Sharing bioimaging data sets and deep-learning models will also be a priority for the Center for Open Bioimage Analysis, an effort funded by the US government and led by Carpenter and Kevin Eliceiri, a bioengineer at the University of Wisconsin–Madison.

Another option, ZeroCostDL4Mic, launched last month. Developed by biophysicist Ricardo Henriques at University College London, ZeroCostDL4Mic makes use of Colab, Google’s free cloud service for AI developers, to provide access to several popular deep-learning microscopy tools, including the one Jacquemet uses to automate cell-nuclei labelling in his films. “Everything you need is installed within a couple of minutes,” Jacquemet explains. With a few mouse clicks, users can use example data to train a neural network to complete the desired task (see ‘Wanted: more data’), then apply that network to their own data – all without needing to code.

Researchers who want to use larger data sets or train more-complex models might need to purchase or access extra computational resources beyond Google’s free service.

By easing the way for biologists with scant know-how and resources to use deep learning, Henriques says, ZeroCostDL4Mic acts like “a gateway drug” for AI, luring researchers to explore the software underlying these tools that will continue to transform research in cancer and beyond.

Esther Landhuis is a science journalist based near San Francisco, California.

1. von Chamier, L. et al. Preprint at bioRxiv <https://doi.org/10.1101/2020.03.20.000133> (2020).
2. Perlman, Z. E. et al. *Science* **306**, 1194–1198 (2004).
3. Ljosa, V. et al. *J. Biomol. Screen.* **18**, 1321–1329 (2013).
4. Ando, D. M., McLean, C. Y. & Berndt, M. Preprint at bioRxiv <https://doi.org/10.1101/161422> (2017).
5. Warchal, S. J., Dawson, J. C. & Carragher, N. O. *SLAS Discov.* **24**, 224–233 (2019).
6. Warchal, S. J. et al. *Bioorg. Med. Chem.* **28**, 115209 (2020).
7. Ma, J. et al. *Nature Meth.* **15**, 290–298 (2018).
8. Ouyang, W. et al. *Nature Meth.* **16**, 1254–1261 (2019).
9. Poirion, O. B., Chaudhary, K., Huang, S. & Garmire, L. X. Preprint at medRxiv <https://doi.org/10.1101/19010082> (2019).
10. Torroja, C. & Sanchez-Cabo, F. *Front. Genet.* **10**, 978 (2019).



Where I work Andrew Digby

Photographed by
Deidre Vercoe/
New Zealand Dept Conserv.

Kakapo are probably the weirdest birds in the world: they're large, flightless, nocturnal parrots. Found only in New Zealand, they are critically endangered and now live only on four predator-free sanctuary islands. Thanks to intensive management of their populations and breeding, we now have 201 birds in our sanctuaries – up from a low of 51 in 1995.

As the conservation biologist on the kakapo team, I work mainly in our office in Invercargill on South Island, New Zealand, 60 kilometres from one of our sanctuary islands. I communicate with international experts about disease, inbreeding and other threats, and about solutions. Last year we had a huge outbreak of aspergillosis, a fungal disease, that made 21 birds ill and killed 9.

Kakapo are unusual in that they manage on very low levels of vitamin D. We're trying to understand this, which might help us to improve supplemental feeding. The birds breed every few years when the rimu tree fruits. The berries have very high levels of

vitamin D, which the birds might need to lay eggs and raise chicks.

During the breeding season, I'll spend months on the sanctuary islands off South Island and near Auckland in the north. We use artificial insemination to boost fertility and avoid inbreeding. We fit every bird with a tracker and an activity monitor, which they wear like a backpack. We use drones to send fresh semen around the island to make the best matches for receptive females.

We're now on near-total lockdown in New Zealand, so we can't monitor the kakapo in person. We'll continue to use our remote-monitoring systems to keep as close an eye on them as we can.

It's rewarding to work with this amazing species that most people never get to see. Sinbad, my favourite (pictured), is inquisitive around people. People fall in love with kakapo when they know them.

Andrew Digby is Science Adviser Kakapo/Takahe for the New Zealand Department of Conservation. **Interview by Amber Dance.**

nature

index

Cancer



IMPROVED PROGNOSIS

Research gains are mounting,
but benefits are unevenly spread

Prevention

A radical
eradication plan

Treatment

The robotic
approach

Survival

Melanoma therapy
shows a way

Survival at all costs

Editorial Catherine Armitage, Bec Crew, Rebecca Dargie, Gemma Conroy, David Payne **Analysis** Bo Wu, Catherine Cheung **Art & design** Madeline Hutchinson, Tanner Maxwell, Wojtek Urbanek **Production** Jason Rayment, Ian Pope, Nick Bruni, Bob Edenbach, Joern Ishikawa **Marketing & PR** Claire Hodge **Sales & partner content** Sabrina Ma, Jennie Xu, Nicole Yu, Pinky Zhang, Alex Yu, Yingying Zhou, Ruffi Lu, Chris Gilloch, Drew Dargis, Jallissa Hamilton, Yuki Fujiwara, Soon Kim, Eri Shimoyama, Ikuko Oba, Shoko Hasegawa, Yoshiko Sugita, Chika Takeda, Natsumi Penberthy **Publishing** Rebecca Jones, Richard Hughes, David Swinbanks.

Nature Index Cancer 2020

Nature Index Cancer 2020, a supplement to *Nature*, is produced by Nature Research, the flagship science portfolio of Springer Nature. This publication is based on data from Nature Index, a Nature Research database with a website maintained and made freely available at natureindex.com.

Nature Editorial Offices

The Campus, 4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0)20 7833 4000
Fax: +44 (0)20 7843 4596/7

Customer services

To advertise with the Nature Index, please visit natureindex.com or email clientservicesfeedback@nature.com. Copyright © 2020 Springer Nature Limited, part of Springer Nature.

All rights reserved.

Some analysts are cynical about the apparent mismatch between spending and outcomes in what's disparagingly called 'the cancer industry'. How can cancer be the second leading cause of death globally, responsible for an estimated 9.6 million deaths in 2018, when a single institution (the US National Institutes of Health, NIH) spent US\$24.4 billion on cancer research in the past four years, not to mention outlays by so many other funders?

As the graph on page S7 shows, researchers are nudging the dial on some types of cancer more than others. This Nature Index supplement focuses on three – cervical, prostate and melanoma – as a lens through which to view the kinds of preventions and treatments that are lengthening survival rates, at least in high-income countries.

Dimensions data provide interesting comparisons on value for money. As a rough indication, looking at the top ten funders' total grants for cancer research from 2010 to 2019 beside their cancer research publications over the same period, the average for the National Natural Science Foundation of China is US\$21,902 per publication. By contrast, for the US National Cancer Institute, part of the NIH and the world's biggest funder of cancer research, it is US\$129,624 per article.

The above analysis is blind to article quality. For that, the indicator is publication in the 82 high-quality journals selected by experts for inclusion in the Nature Index, which, it should be noted, does not include clinical sciences journals. In cancer, as in every other field, China's rise is striking. Its cancer research in the Nature Index rose by an estimated 114.9% from 2015 to 2019, according to our key metric, Share, a fractional count of the proportion of the country's affiliated authors on each article. The number of cancer research articles published in Nature Index journals, identified through a search using Dimensions, grew by 25.8% over the same period, more than four times the growth for articles overall. One reason that cancer outcomes seem not to be improving in line with research output is that improved treatments have not been accessible to all, as our stories about cervical and prostate cancer explain (pages S2 and S5). Therapies, let alone the latest treatments, may be out of reach in the low- and middle-income countries where 70% of global cancer deaths occur. That's not a problem science alone can solve.

Catherine Armitage
Chief editor

**On the cover**

Artistic rendition of a breast cancer cell disintegrating, from an image by Anne Weston, Francis Crick Institute.

Contents

- S2 A global drive towards elimination**
The push to eradicate cervical cancer.
- S5 A closer look at a revered robot**
Widespread preference for robotic surgery comes at a cost.
- S14 Cellular roadblocks trip melanoma hijackers**
A game-changing class of immunotherapy drugs.
- S24 Coming at cancer from all angles**
The researchers making a difference.
- S30 The tables**
How the world's institutions stack up in cancer research.



Akello Faith receives her HPV vaccine at a mobile health clinic in Ochaga, Uganda.



A global drive towards elimination

Vaccination is picking up where screening left off in reducing cervical cancer rates. **By Sarah DeWeerd**

Cervical cancer is already a prevention success story. The widespread use of the Pap smear, a screening test to identify pre-cancerous and cancerous cells collected from the cervix, has led to steep declines in diagnoses and deaths from cervical cancer over the past half-century, especially in high-income countries. Cervical cancer is still one of the most common cancers in women, so the World Health Organization (WHO) has an ambitious goal: to eliminate it as a public health problem. Its World Health Assembly plans to vote on a strategy to put all countries on a path to cervical cancer elimination by 2030.

Nearly all cases of cervical cancer are caused by the human papilloma virus (HPV), which is spread through sexual contact. Most sexually active adults have been exposed, although only some strains cause cervical cancer.

Since 2006, the HPV vaccine has achieved further stunning results in cervical cancer prevention. For example, in England, where a national vaccination programme for adolescent girls began in 2008 and expanded to include boys last year, public health officials revealed in January 2020 that less than 2% of sexually active young women tested in 2014 to 2018 carried HPV strains 16 and 18. Those strains significantly increase the risk of cervical cancer and are responsible for 70% of cases. The results are “very impressive”, says Lois Ramondetta, a gynaecologic oncologist at the MD Anderson Cancer Center in Houston, Texas, which is ranked fourth in the Nature Index for cancer research output. “It’s more evidence for how effective the vaccine is, as well as for the important concept of herd immunity.” Herd immunity is the idea that high vaccination rates can virtually eliminate the virus at a population level.

In cancer prevention and women’s health circles, the news raises hopes that the WHO’s goal is in sight and cervical cancer could soon become vanishingly rare.

A new screening method

To epidemiologists and public health officials, ‘eliminate’ has a specific definition: fewer than 4 new cases per 100,000 women per year. The current global rate stands at 13.1 per 100,000;

more than 80% of diagnoses and deaths occur in low- and middle-income countries. Reaching that benchmark depends on both vaccinating adolescent girls and screening adult women. But as far as screening goes, the Pap tests that have been key to reducing the burden of cervical cancer may play a relatively small role going forward. Testing cervical samples for the presence of DNA from high-risk HPV strains, rather than for abnormal cells, as in a Pap smear, is now seen as a more effective way to identify women at risk of developing cervical cancer, partly because it may identify women at risk earlier than a Pap smear can, and is more objective. HPV testing replaced the Pap smear in Australia in December 2017 and the US Preventive Services Task Force recommended the test, alone or in combination with Pap smears, in 2018. Many women in low- and middle-income countries lack access to any form of cervical cancer screening, but public health researchers are investigating how to roll out HPV testing programmes in these settings.

Success story

HPV testing of adult women is crucial to reducing cervical cancer rates in the short term, says Karen Canfell, director of research at the Cancer Council of New South Wales in Australia, because cervical cancer typically develops 15 to 20 years after HPV exposure in healthy women. In England, 96% of patients diagnosed with early stage cervical cancer are still alive one year later, compared with only 50% of those with a late-stage diagnosis.

But in its early stages cervical cancer’s symptoms are often non-specific, such as unusual vaginal bleeding, or absent altogether, making screening tests all the more important. Vaccination, meanwhile, is key to driving down cancer rates over the longer term.

Rapid scale-up of both vaccination and HPV screening could prevent up to 13.4 million cases of cervical cancer worldwide in the next 50 years, and enable all countries to reduce incidence below 4 per 100,000 by the end of the century, according to a global modelling study led by Canfell.

Australia could be the first to reach the WHO’s benchmark for cervical cancer elimination. Its

nationwide, publicly funded, school-based HPV vaccination programme was introduced for girls in 2007, and expanded to include boys in 2013. The country's universal health-care system provides access to cervical cancer screening at no or low cost to most women.

If high rates of both vaccination and screening continue, Australia could eliminate cervical cancer as early as 2028, according to calculations by Canfell and her colleagues.

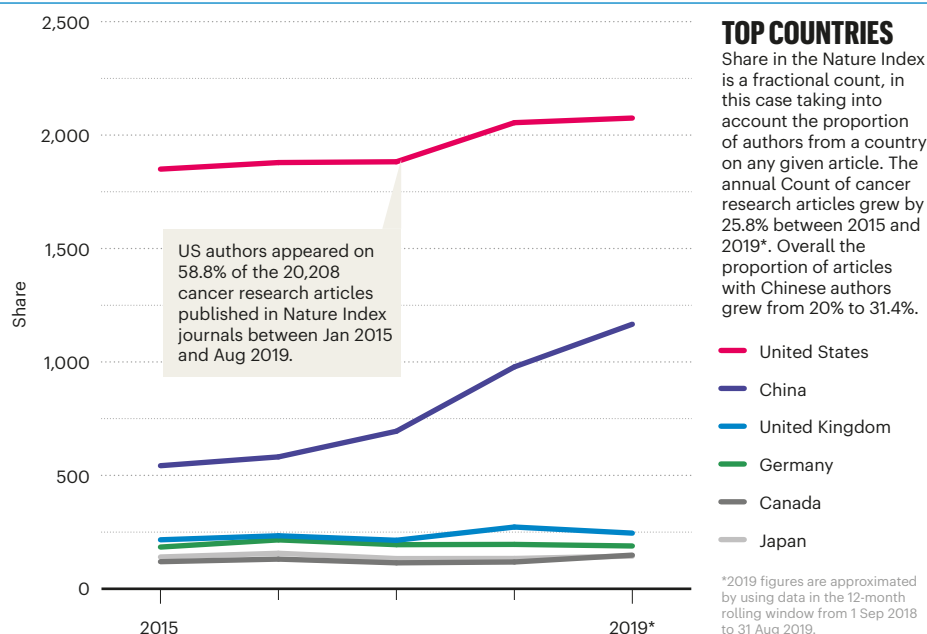
But reaching the benchmark of elimination in Australia and elsewhere does not mean inequities will be removed, she cautions. Cervical cancer rates are twice as high, and deaths three times as frequent, among Aboriginal women in Australia compared to non-Aboriginal women. According to the US Centers for Disease Control and Prevention, cervical cancer rates are higher in some populations, including black and Hispanic women, as well as women living in the south, than in the United States as a whole.

Strategies for better coverage

Reaching the WHO's goal will also be challenging in low- and middle-income countries such as India and Kenya, where both cost and poor health-care infrastructure pose a barrier. Cervical cancer rates are relatively high in many of these places because Pap testing has not really filtered down, says Surendranath Shastri from the MD Anderson Cancer Center.

Shastri has worked on studies of several alternative approaches to cervical cancer screening in India, where laboratory facilities and trained pathologists who can interpret Pap smears are scarce outside large cities. He and his colleagues have shown that a test known as VIA, in which the cervix is coated with acetic acid and then inspected for white areas that may signify cancer or precancerous lesions, can be performed by community health workers and reduces cervical cancer mortality.

Self-sampling, in which women swab their own cervical tissue for HPV testing, could help to increase access to cervical cancer screening in areas where women's health specialists and facilities for pelvic exams are inadequate. "Being able to offer women a self-collected test has opened up a lot more options for when and where they can be screened," says Megan Huchko, a gynaecologist and director of the Center for Global Reproductive Health at Duke University in Durham, North Carolina, ranked 28th in the Nature



Index for cancer research. For example, Huchko and her colleagues working in East Africa have had success in offering the screening to women at community health fairs and door-to-door visits by health-care workers; women often receive their results by text message. These strategies can increase screening coverage from 5% to as much as 70%, they have found. However, the low-cost HPV tests that are currently available don't pinpoint the highest risk strains of the virus, so some women may be receiving unnecessary treatment for medium-risk strains, Huchko says.

One hundred countries have added the HPV vaccination to their recommended vaccine schedules since 2006, but most of these are high-income countries, and altogether they cover only 30% of the global population of girls who need the vaccine. This number is slowly climbing: both Kenya and Uzbekistan launched HPV vaccination programmes in late 2019.

In high-income countries, the HPV vaccine can cost as much as US\$100 per dose (at least two doses are needed for full protection). Gavi, a global public-private partnership to increase vaccine accessibility, has helped 27 countries access the vaccine for as little as \$4.50 per dose. The advent of HPV vaccines produced more cheaply in middle-income countries such as China and India and shipped around the world will also reduce costs – the first such vaccine

was approved in China at the end of 2019. But, Shastri cautions, "It's not just about providing the vaccines. There are a whole lot of logistics that go into the vaccine delivery."

Some high-income countries have also struggled to attain the cervical cancer screening and HPV vaccination rates that will be necessary to eliminate the disease. In the United States, a fragmented health-care system limits access to screening for many women. And only about half of adolescents aged 13 to 17 were up to date on their HPV vaccination in 2018.

Vaccine hesitancy is common in the United States, and it hasn't helped that the HPV vaccine is associated with the issue of young women's sexuality. The vaccine was originally advertised as an anti-STD for girls rather than an anti-cancer vaccine, Ramondetta says. That led to reluctance by government officials to mandate the vaccine, especially in conservative states, out of fear that it would promote promiscuity. In addition, gynaecologists were initially responsible for delivering the vaccine, when most girls of the appropriate age would still be cared for by paediatricians. Still, Ramondetta says, vaccination rates in the United States "continue to go up every year, which is pretty amazing".

Sarah DeWeerd is a science writer in Seattle, Washington.

TOWARDS ELIMINATION TIMELINE OF ACTION

2013

Australia's cervical cancer vaccination programme expands to include boys.

2014

The US Food and Drug Administration approves vaccine against 9 HPV strains.

2014

The WHO updates guidelines on cervical cancer prevention and control.

2018

The WHO calls for elimination of cervical cancer as a global public health problem.

2019

Kenya and Uzbekistan launch HPV vaccination programmes.



KEN LEANFORD FOR NATURE

Urology fellow, Jeremy Fallot, and nurse, Shauna Harnedy, assist in robotic surgery by Ruban Thanigasalam (out of view) in Sydney, Australia.

A closer look at a revered robot

The da Vinci robotic system has become the ubiquitous method for prostate removal, but its high cost is raising questions. **By Bec Crew**

Loved by surgeons and patients alike for its ease of use and faster recovery times, the da Vinci surgical robot is less invasive than conventional procedures, and lacks the awkwardness of laparoscopic (keyhole) surgery. But the robot's US\$2-million price tag and negligible effect on cancer outcomes is sparking concern that it's crowding out more affordable treatments.

There are more than 5,500 da Vinci robots globally, manufactured by California-based tech giant, Intuitive. The system is used in a range of surgical procedures, but its biggest impact has been in urology, where it has a market monopoly on robot-assisted radical prostatectomies (RARP), the removal of the prostate and surrounding tissues to treat localized cancer. Uptake in the United

States, Europe, Australia, China and Japan for performing this procedure has been rapid. In 2003, less than 1% of surgeons in the US performed a RARP in preference to open or laparoscopic surgery. By 2014, RARP accounted for up to 90% of radical prostatectomies across the country. When it comes to prostate cancer surgery in the United States, says Benjamin Davies, surgeon and professor of

urology at the University of Pittsburgh, “the die is cast; there is only robotic surgery”.

After lung cancer, prostate cancer is the second most common cancer in men worldwide. It affects the walnut-sized prostate gland, which sits up against the urethra, between the rectum and bladder, and secretes prostate fluid, a component of semen. The prostate’s proximity to the blood vessels, muscles and a fragile web of nerve bundles that control erectile and bladder function, demands extreme surgical precision in its removal, a procedure that is generally recommended if the disease has not yet spread. Whereas an open patient needs to be cut from naval to pubic bone in order to access the prostate, a robot-assisted procedure requires a few small abdominal incisions.

Known as a master–slave system, the da Vinci comprises three main components. The tower (or ‘slave’) wields three arms equipped with instruments such as forceps, hooks and needle-drivers, and a fourth holds cameras capable of 15 times magnification. The console (‘master’) is where the surgeon sits, a few metres from the patient, remotely operating the robot arms while watching through a 3D stereoscopic monitor. A separate cart contains image-processing equipment.

Surgeons prefer to use the da Vinci robot because it offers improved visualization and hand and wrist flexibility, and they can be seated throughout the 2- to 4-hour procedure. “We can see the anatomy of the prostate like we have never seen it before,” says Freddie Hamdy, Nuffield professor of surgery and urology at the University of Oxford, UK, which is ranked 26th in the Nature Index for cancer research output.

Cancer outcomes equal

Whether these improvements translate to better long-term outcomes for the patient, however, remains unclear. Ruban Thanigasalam, associate professor of robotic surgery at the University of Sydney and clinical lead



Success of robotic surgery lies solely in the skill of a surgeon, says Ruban Thanigasalam.

in prostate cancer research at the Institute of Academic Surgery in Australia, is conducting a trial comparing open and robotic surgery. The preliminary results support what has been widely accepted by surgeons for years: robotic-surgery patients experience reduced blood loss, less pain and shorter recovery time, but the longer-term outcomes are equivalent.

“Anecdotally, we find that recovery of continence is earlier in the robotic group, but after 12 months, there is no major difference between the two for urinary control and sexual function,” says Thanigasalam. For the cancer itself, he adds, the outcomes are the same.

“Several international studies looking at tens of thousands of patients have all shown that there is absolutely no difference in cancer outcomes between robotic and open surgery.”

Thanigasalam stresses that the outcomes of robotic surgery remain dependent on the surgeon’s skills, a sentiment echoed by Davies: “It’s always the surgeon’s hands, not the technology we use.”

Even da Vinci’s proponents acknowledge the temptation to overplay its ability. “We all love a good robot,” says Richard Sullivan, professor of cancer and global health at King’s College London and director of the Institute of Cancer Policy in the United Kingdom. “Human beings, particularly surgeons, are incredibly neophilic. We love this sort of thing, it gives us authority. And the patient will think that, because you’ve got all of this fancy kit, you must have better outcomes. But that’s not true, the robot is not an indicator of quality.”

Accessibility gap

According to a 2017 report by the Royal Australasian College of Surgeons and Australian health insurance provider, Medibank, the cost of a prostate cancer procedure varied nationwide from Aus\$14,553 to Aus\$55,928 (US\$9,165 to US\$35,222). The use of robotics, the report states, “can substantially increase the cost”.

Despite questions over value for money, business is booming. In 2018, the global surgical robots market was worth US\$6.8 billion, and it’s predicted to hit \$17 billion by 2025. In response to the surge in robotic surgery, the US Food and Drug Administration (FDA) urged patients and health-care providers to exercise caution last year, particularly with regards to breast and cervical cancer, citing a lack of long-term evidence. “The problem is, once it becomes adopted, it can be very difficult to pedal it back,” says Hamdy.

A “massive inequality gap” is opening between hospitals that can afford the robot, and those that can’t, says Sullivan. “In many countries, we’re fighting for patients because of choice and competition. If I’ve got a robot, I can sell that fact to patients, and they’ll come to me rather than the centre down the road.”

RISE OF THE ROBOT A STEADY HAND-OVER

1982

Patrick Walsh from Johns Hopkins University performs the first nerve-sparing radical prostatectomy, making it possible to preserve sexual function and urinary continence in some patients.

1995

Intuitive, da Vinci’s manufacturer, is founded by surgeon Frederic Moll, engineer Robert Younge, and venture capitalist John Freund.

1998

The first commercial sale of a da Vinci robotic system is made to the Leipzig Heart Center in Germany.

2000

The da Vinci is the first robotic system to gain FDA approval for general laparoscopic surgery.

2001

An account of the first robotically assisted radical prostatectomy, which was performed using a da Vinci system, is published in the *BJU International* by J. Binder and W. Kramer at the Johann Wolfgang Goethe University in Germany (Binder, J. & Kramer, W. *BJU Int.* **87**, 408–410; 2001).

A 2019 paper co-authored by Sullivan for the World Health Organization found that competition between hospitals with and without surgical robots contributed to the closure of 25% of radical prostatectomy centres in the English National Health Service. This focus on “expensive medicines for wealthy patients in wealthy countries”, the paper states, is putting low-income groups at a disadvantage by crowding out spending on the development of preventative measures (R. Sullivan and A. Aggarwal in *Reducing Social Inequalities in Cancer: Evidence and Priorities for Research*, IARC Monograph, 2019).

A stark divide also exists between high-income and low- and middle-income countries, which makes it difficult to treat patients across borders, says Sullivan. “Most of our juniors [in the UK] have been trained in minimally invasive and robotics surgery,” he says. “They’re saying, ‘If I want to work somewhere like Zambia or India, I’m screwed if I’ve only done minimally invasive or robotics.’ Outside the high-income settings, these services aren’t available.”

Competition could drive the price of the da Vinci robot down, such as from UK-based CMR Surgical, which has raised \$240 million since 2016 for its Versius robot, and Verb Surgical, a partnership between Johnson & Johnson and Alphabet.

Improved screening could see fewer men undergoing surgery in the first place. There is evidence that the benefits of the prostate-specific antigen (PSA) blood test, which, along with a digital rectal examination, is the most common way to screen for prostate cancer, may not outweigh the potential harm of misdiagnosis leading to unnecessary surgery or radiation. Researchers from the Queen Mary University of London and University of East Anglia, UK, are developing blood and urine tests to be used in conjunction with the PSA.

Bec Crew is a senior editor at Nature Index.

2009

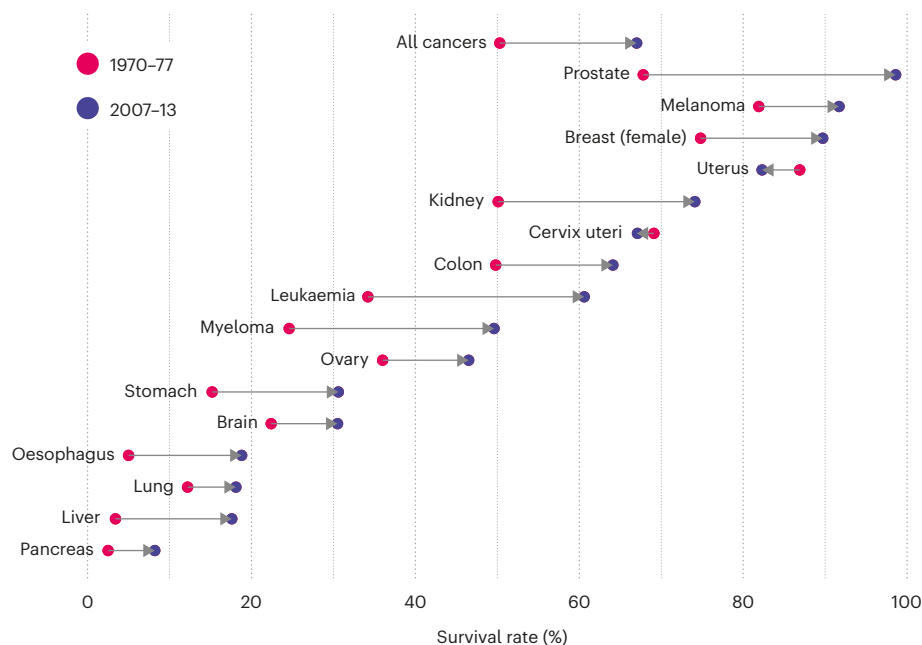
86% of prostate cancer surgeries in the United State are robot-assisted operations.

2019

Intuitive’s stock price grows 66% from US\$312 in 2017 to \$520 in 2019. Its total revenues grow from \$3.7 billion in 2018 to \$4.5 billion in 2019 (preliminary).

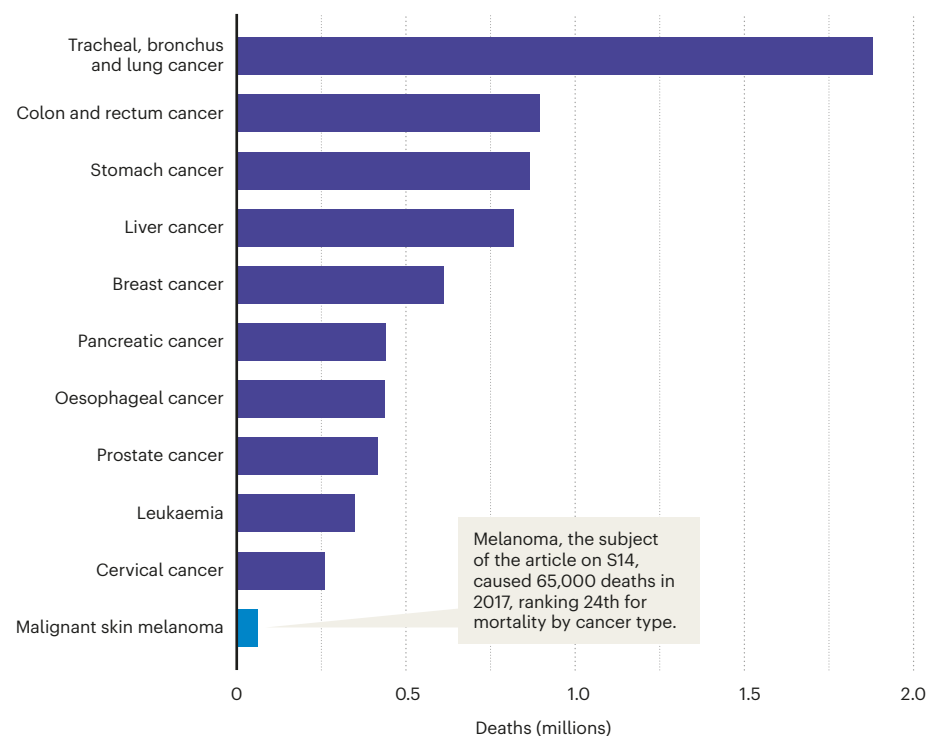
IMPROVED OUTCOMES FOR MANY CANCER TYPES

Changes in average five-year survival rates in common cancer types in the United States since the early 1970s. The five-year survival rate is the percentage of people who live longer than five years after diagnosis. The cost of cancer diagnosis and treatment globally rose 27% to US\$107 billion between 2010 and 2015, according to the IMS Institute for Healthcare Informatics. It is projected to rise again by 40% to \$150 billion in 2020 in constant dollar terms.



WHERE CANCER STRIKES

Deaths from cancers globally in 2017, when there were 9.56 million cancer deaths overall. Lung cancer and breast cancer are the two most common types but lung cancer was the biggest cause of cancer death that year while breast cancer was 5th. Colorectal cancer is the 3rd most common type and was the 2nd biggest cause of cancer deaths in 2017.



SOURCE: INTERNATIONAL HEALTH METRICS AND EVALUATION, GLOBAL BURDEN OF DISEASE FROM ROSER, M. & RITCHIE, H. "CANCER" [HTTPS://OURWORLDINDATA.ORG](https://ourworldindata.org) (2020)

Cellular roadblocks trip melanoma hijackers

Success for a class of immunotherapy drugs is changing the face of treatment. **By Bianca Nogrady**

When Jedd Wolchok began working in the area of melanoma 20 years ago, the average life expectancy for a patient with advanced disease was six or seven months.

Now his waiting room is full of people coming back for their third or fourth year of follow-up, sharing their stories of survival with the newly diagnosed, giving hope where just a decade ago there was little.

“That gives you a sense of the human impact of this,” says Wolchok, a medical oncologist and director of the Parker Institute for Cancer Immunotherapy at the Memorial Sloan Kettering Cancer Center in New York, ranked fifth in the Nature Index for cancer research output.

Transformative treatment

Behind this transformation in melanoma survival rates is a class of drugs called checkpoint inhibitors, the first of which was approved nine years ago. Checkpoint inhibitors are a form of cancer immunotherapy – treatments that stimulate the immune response to cancer cells. Checkpoint inhibitors are not the first form of cancer immunotherapy, but they are, so far, among the most successful, particularly in melanoma. They’re also having a big impact in lung and urinary tract cancers. “Melanoma is the most sensitive type of cancer to checkpoint inhibitors,” says James Larkin, medical oncologist at the Royal Marsden Hospital in London.

“Some patients who were quite sick were improving really, really quickly, which we’d never seen before.”

But no one is sure why. Some patients respond well to checkpoint inhibitors, but others don’t respond at all, for reasons that are also not yet understood.

Checkpoint inhibitors work by preventing tumour cells from hijacking, and therefore avoiding, the cellular immune response that should eliminate them. Their discovery came about in the late 1990s, when two groups of

researchers from the United States and Japan uncovered a series of interactions between cell-surface receptors and proteins that led to the death of immune T cells.

T cells are the cells that would normally lead the charge against cancer and other threats. They have a receptor on their surface called PD-1 (programmed cell death protein 1). When that receptor is engaged, it triggers the T cell to rupture – one of the many checkpoints that have evolved to help keep the immune system from over-reacting.

The protein that engages that receptor is PD-L1 (PD ligand 1). It turns out that many human cancers also produce PD-L1, the factor that tumours are using to hijack the checkpoint and engage the T-cell death receptor to stop the response against them.

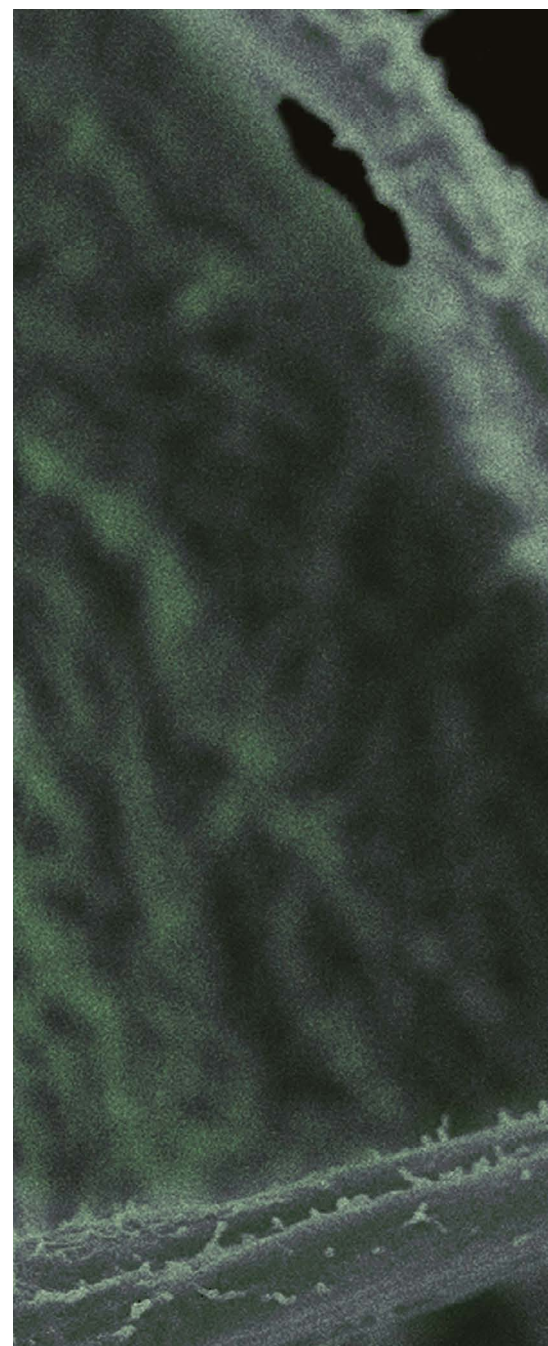
Scientists showed that inhibiting this tumour-hijacked checkpoint could unleash an immune response against the tumour.

A sense of possibility

The first checkpoint inhibitor drug, ipilimumab, was approved by the US Food and Drug Administration in March 2011 for the treatment of melanoma that had spread or that could not be treated surgically. Compared with a melanoma vaccine, itself a new therapeutic approach being trialled, the drug significantly improved survival rates. Although it worked in only around one in five patients, the benefits in those patients were dramatic, Larkin says. “We really had a sense then of the possibilities.”

Ipilimumab was followed by pembrolizumab in September 2014, and nivolumab just three months later. All of these, and newer checkpoint inhibitors, are now in widespread use, although they’re expensive for patients, particularly in countries without public health insurance schemes. A course of intravenous checkpoint inhibitor therapy can cost US\$150,000–250,000 per year.

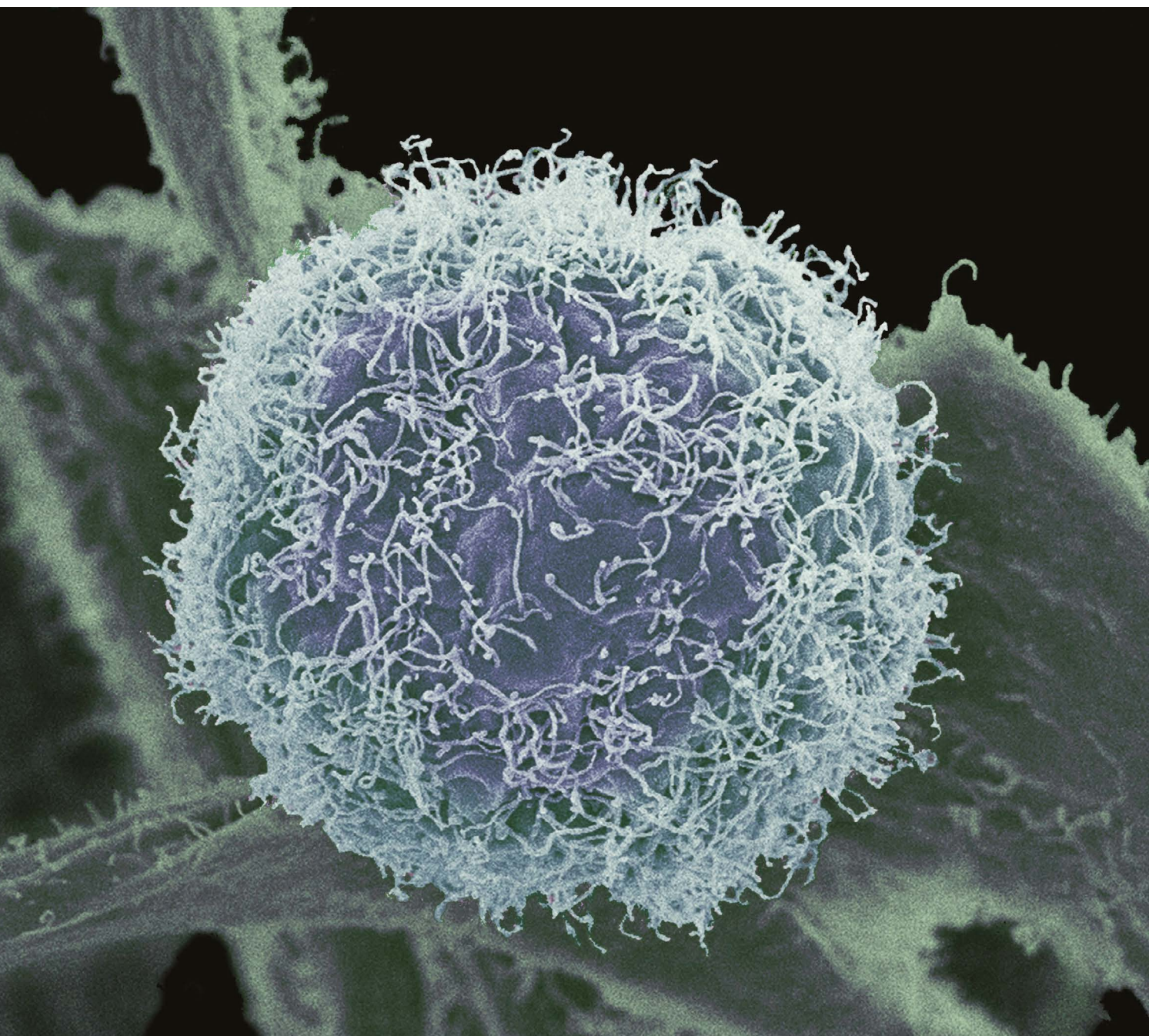
The most spectacular results so far with checkpoint inhibitor therapy have come from trials combining two different checkpoint inhibitors, such as ipilimumab and nivolumab. Larkin and Wolchok were both involved in the CheckMate 067 study, which began in



July 2013 and compared ipilimumab alone with nivolumab alone, and with ipilimumab plus nivolumab in 945 people with advanced untreated melanoma.

“It was a blinded trial, so you didn’t know which treatment the patients were getting,” Larkin says. “And it was really striking that some patients who had symptoms or were quite sick were improving really, really quickly, which we’d never seen before.”

The combination was so successful that a paper published in the *New England Journal*



ANNE WESTON, FRANCIS CRICK INSTITUTE

The treatment of many melanomas is now starting with immune checkpoint inhibitors, rather than major surgery.

of Medicine in late 2019 showed that 52% of patients were alive after five years, compared with 44% of patients on nivolumab alone and 26% of patients on ipilimumab alone (J. Larkin *et al. N. Engl. J. Med.* **381**, 1535–1546; 2019). As often with clinical trials, checkpoint inhibitors were first tested in the most severely affected patients, those whose cancer was untreatable with surgery or which had spread despite existing treatments. But with each new trial showing unprecedented survival rates, questions would arise as to whether these drugs should

be used earlier in the disease, even before it had spread.

Grant McArthur, a medical oncologist and head of the molecular oncology laboratory at the Peter MacCallum Cancer Centre in Melbourne, Australia, says checkpoint inhibitors have brought a paradigm shift in the management of melanoma. “We see patients, who previously would have had large, complex surgical procedures that are associated with substantial morbidity, who now will start with the immune checkpoint inhibitors,” he says.

“The idea that immunotherapy could replace surgery is being entertained for the first time.”

It’s not all good news. Checkpoint inhibitors come with some potentially serious side effects, many as a result of an over-active immune response, which is linked to inflammation in the bowel, lung, heart, skin and other organs. And around half of the patients with advanced disease don’t respond as spectacularly, or at all, to checkpoint inhibitors.

Some survive longer than they might have done without treatment, or have a longer

period until their disease progresses. However, the CheckMate 067 study found that 48% of patients had died within five years, despite treatment with a combination of checkpoint inhibitors. There's palpable frustration over why no one can explain this. It's an active area of research, and there are early suggestions about what might be the deciding factors. One clue is that people who seem to get the most benefit from checkpoint inhibitors are those whose immune systems are already putting up a fight when they start treatment, says Wolchok.

"The best evidence for that comes from pathology studies, which have shown that tumours that already have T cells in them are the ones where you see responses," he says. "What the checkpoint inhibitors are doing in general is allowing a pre-existing immune response to become more effective."

There's also evidence that patients with cancers caused by a certain genetic condition called mismatch repair deficiency may actually respond better to checkpoint inhibitors, regardless of their cancer type.

Into the unknown

Another feature that seems to be linked to better response rates is what's called the mutation burden of the tumour, the number of genetic mutations present in the genome of an individual's cancer. Just as exposure to cigarette smoke causes the mutations that are common to lung cancers, exposure to ultraviolet radiation causes a set of mutations that are common features of skin cancer. But individuals with skin cancer that grows in parts of the body that are less exposed to the sun may have a lower mutation burden, and that seems to make them less likely to respond to checkpoint inhibitors.

"The hypothesis is that cancers that have a lot of mutations have many abnormal-appearing proteins, which makes them look different from the normal cell that they came from," says Wolchok. "That is something that the immune system at baseline is able to survey for."

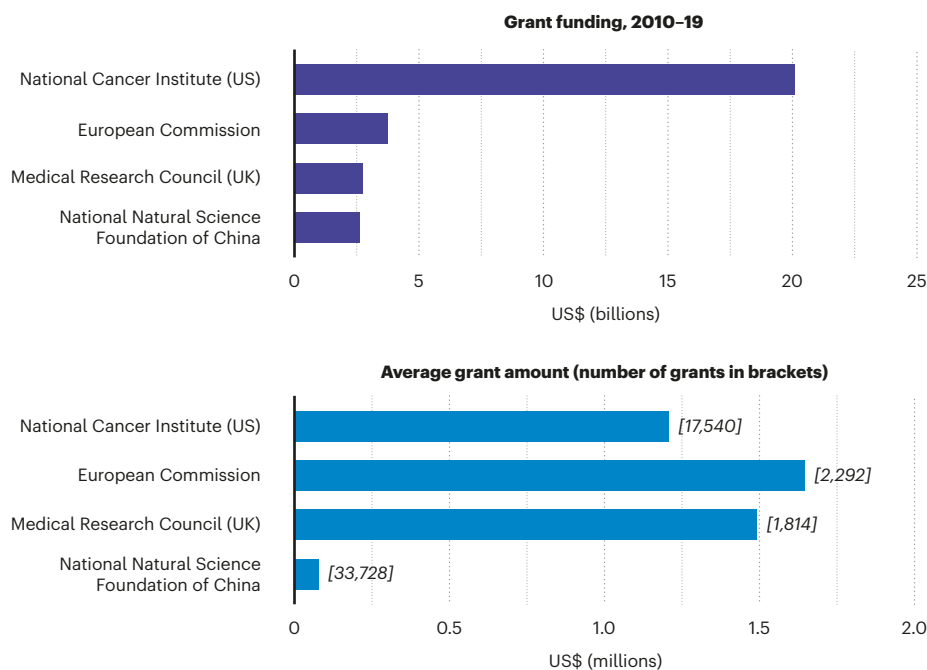
Given the survival rates among people who do respond to checkpoint inhibitors, is it time to start talking about a cure for melanoma? Oncologists are wary of the word, preferring to talk about long-term survivorship, which is itself a novel concept in melanoma.

"If you've no longer got a disease that 20 years ago had a survival of six to nine months, and it turns out that you're a long-term survivor, what does that look like?" Larkin asks. "Curing metastatic solid tumours isn't something that we've ever really faced before."

Bianca Nogrady is a science writer in the Blue Mountains, Australia.

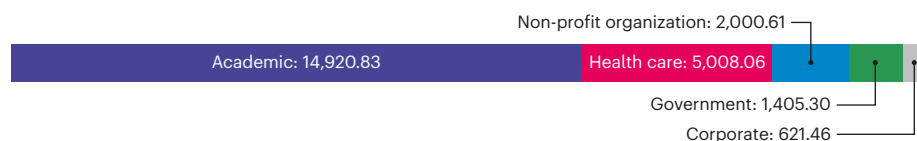
SELECTED TOP FUNDERS OF CANCER RESEARCH

Although the US National Cancer Institute (NCI) is by far the world's biggest funder of cancer research, China's National Natural Science Foundation makes many more individual grants of much smaller amounts, and the European Commission's average grant size is 40% larger than the NCI's.



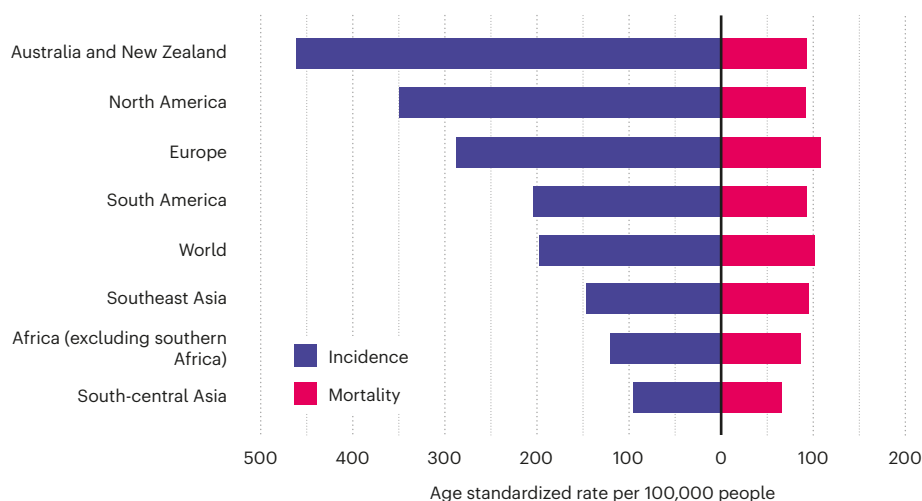
RESEARCH BY SECTOR

Sectoral contributions to cancer articles over the period according to their Nature Index Share, a fractional count that takes into account the proportion of authors from that sector on any given article. Total Share for cancer articles across all sectors was 20,208; the figure for total Share by sector is larger because some institutions belong to multiple sectors.



LIVING AND DYING

Incidence and mortality rates in different regions. Accuracy may vary owing to limited data quality and coverage, particularly in low- and middle-income countries. According to the World Health Organization, approximately 70% of deaths from cancer occur in those countries.



Coming at cancer from all angles

The search for disease mechanisms and treatments is one of the biggest collaborative efforts in science. These researchers are significant contributors.



Biostatistician Heidi Kosiorek works at the crossroads of maths and medicine, sifting through data to weigh the best course of action based on cancer types and patient profiles.

“It’s impossible to do cancer research on your own if you want to do something that makes sense for the disease,” says lung cancer researcher, Niki Karachaliou, referring to the diverse teams of physicians, clinicians and other researchers whose different perspectives help create a shared understanding of this complex disease.

Karachaliou and two other researchers whose collaboration networks are shown here were selected for the strength of their publication count in Dimensions. They were drawn from an elite group of researchers in cancer, who were authors in Nature Index between 2015 and 2019 and whose first authorship on an article in Dimensions dates between 2010 and 2014. Authors from the United States dominate the group, as might be expected, given the US leads the field. Although the closest connections for two of the collaboration networks shown here are domestic, the third reflects activity between the United States and China, the top two countries for cancer research.

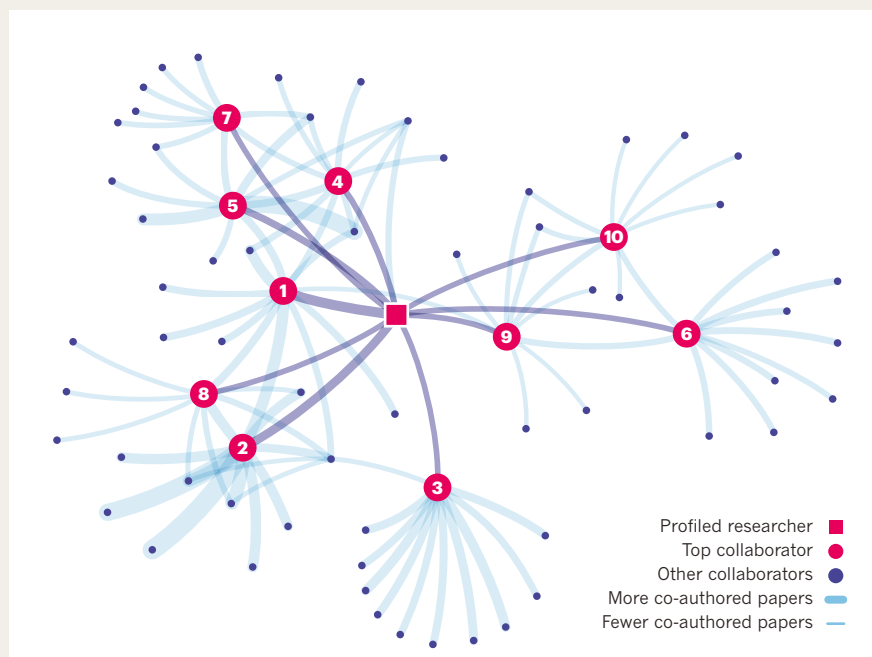
Among the top five countries for cancer research in the index, the United States and China are the most self-sufficient, with internationally collaborative articles on cancer comprising only 51.8 and 49.8%, respectively, of their total cancer articles. For the third-ranked country, the United Kingdom, 81.6% of cancer articles are internationally collaborative, and for Germany, fourth, 80.7% of cancer articles share authorship with researchers outside Germany.

Number cruncher: A biostatistician leverages her maths expertise to improve cancer care
Heidi Kosiorek
Annual average publications count: 30.2

Predicting how various treatments could affect an individual cancer patient requires more than a medical understanding of the genetics and molecular mechanisms underlying the disease. It involves sifting through

WELL-CONNECTED

The graphs show each profiled researcher's top ten research collaborators, ranked by the number of papers they have co-authored, and their collaborators' top ten co-authors. In this series of graphs, each node represents a researcher, with line widths sized by the strength of the collaboration. Lines connect the profiled researcher (square) to their co-authors (numbered nodes). Not all collaborators mentioned in the text will be reflected in these networks, depending on the number of resulting papers to date.



TOP COLLABORATORS FOR HEIDI KOSIOREK

- Amylou Dueck**
Mayo Clinic, United States
- Ruben Mesa**
The University of Texas MD Anderson Cancer Center, United States
- Jeanne Palmer**
Mayo Clinic, United States
- Donald Northfelt**
Mayo Clinic, United States
- Barbara Pockaj**
Mayo Clinic, United States
- Curtiss Cook**
Mayo Clinic, United States
- Karen Anderson**
Arizona State University, United States
- Robyn Scherber**
The University of Texas MD Anderson Cancer Center, United States
- Nina Karlin**
Mayo Clinic, United States
- Patricia Verona**
Mayo Clinic, United States

reams of data to identify, by weight of numbers, the tumour and patient characteristics that could influence success or failure. It's about separating the trends from the flukes, the biomarkers from the outliers.

This first occurred to Heidi Kosiorek in the mid-1990s, during an internship working as a research assistant in the emergency department of an Ohio hospital. "I think it's even more true today" at a time when personalized and precision medicine have become buzzwords, says Kosiorek, now a biostatistician at the Mayo Clinic's Scottsdale, Arizona, campus.

Kosiorek had planned to go to medical school, but with her eyes opened to the intersection between mathematics, which had long been her strength, and medicine, she pursued biostatistics instead. That decision kicked off a prolific career. Since 2015, Kosiorek has authored or co-authored an average of 30 publications per year. For more than a decade, she worked at the University Hospitals Case Medical Center in Cleveland on studies of ovarian, endometrial and cervical cancers. Since joining the Mayo Clinic in late 2014, she's developed expertise in breast cancer, and often teams up with Mayo Clinic colleagues, Barbara Pockaj, a surgeon, and Donald Northfelt,

a medical oncologist, to investigate topics such as tumour genetics and detecting recurrence.

The Mayo Clinic receives nearly US\$120 million in annual grant funding for cancer research, keeping biostatisticians at its three major campuses busy. Choosing which projects to work on "is a challenge, for sure", says Kosiorek, "because I want to do it all".

As an assistant supervisor, Kosiorek helps assign roughly 20 statisticians and statistical programmers to projects and oversees their work. Her mentorship of many junior scientists partly explains her impressive publication rate. "You end up being a part of more projects because of that," she says.

Kosiorek's collaborations extend beyond the Mayo Clinic. She is the lead biostatistician for the Myeloproliferative Neoplasms Research Consortium, a US National Cancer Institute-funded initiative to improve treatment for a rare chronic blood cancer that sometimes develops into leukaemia. She works closely with researchers from several institutions across the United States and Canada that are part of the consortium.

A crucial part of individualized medicine is taking into account the patient's preferences regarding outcomes and quality-of-life

impacts. With that in mind, Kosiorek and Pockaj are running a study on follow-up surgeries after breast reconstruction. They want to categorize the kinds of procedures that are needed for cosmetic considerations, for example, and how often, so that breast-cancer patients can make more informed decisions when considering reconstruction after a mastectomy.

"What drives me is helping physicians find what's best for their patients through the data," Kosiorek says. "There are many days where it doesn't really feel like work." **Sarah DeWeerd**

Biomedical naturalist: A biomedical engineer harnesses an organic drug delivery system – the body's own cells

Chao Wang

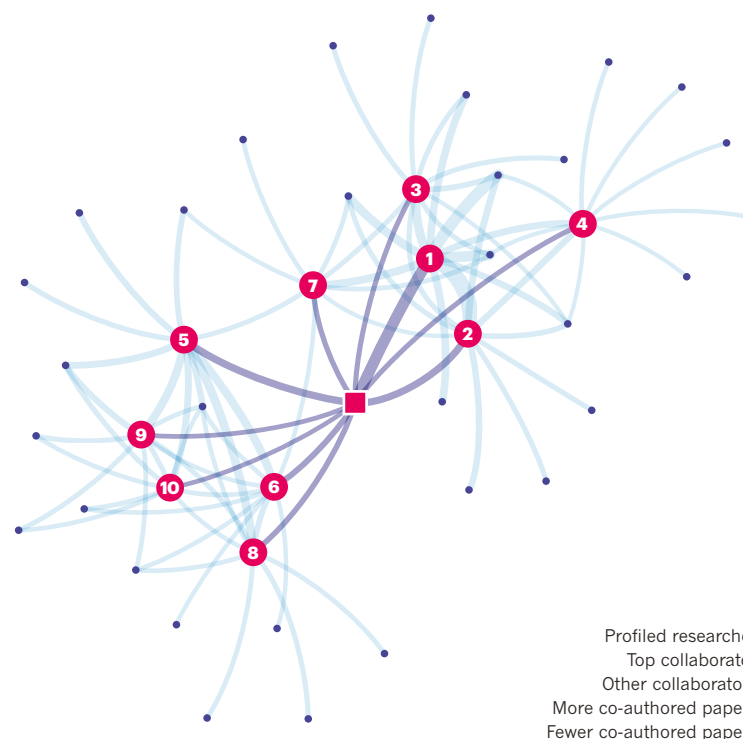
Annual average publications count: 13.4

Immunotherapy, which harnesses the immune system to attack tumour cells, is one of the hottest fields in cancer research. One approach involves attaching immune-boosting drugs to nanoparticles made of gold or iron oxide and injecting them into the patient.

In his pursuit of new cancer-fighting drugs, biomedical engineer Chao Wang eschews such

TOP COLLABORATORS FOR CHAO WANG

1. **Zhuang Liu**
Soochow University, China
2. **Liang Cheng**
Soochow University, China
3. **Chao Liang**
Soochow University, China
4. **Yonggang Li**
The First Affiliated Hospital of Soochow University, China
5. **Zhen Gu**
University of California, Los Angeles, United States
6. **Quanyin Hu**
Massachusetts Institute of Technology, United States
7. **Qian Chen**
Soochow University, China
8. **Wujin Sun**
University of California, Los Angeles, United States
9. **Jicheng Yu**
The University of North Carolina at Chapel Hill, United States
10. **Jinqiang Wang**
University of California, Los Angeles, United States



NATURE INDEX/DIMENSIONS FROM DIGITAL SCIENCE

synthetic materials and looks instead to more organic drug carriers — the body's own cells. "Nature is the best engineer," he says.

In addition to being expensive to produce, metallic nanoparticles could potentially be toxic over the long term, says Wang. He's investigating the humble red blood cell, the most abundant cell type in the human body, as a safer alternative. "Red blood cells may be the ideal carrier," he says.



Chao Wang.

In 2019, Wang and his team described how immune-stimulating molecules called antigens were administered to mice using their own red blood cells. The cells were extracted, modified with antigens, and then re-injected, where they were taken up by the spleen. As reported in *Science Advances*, the treatment spurred the immune systems of the mice, which helped to slow the tumour growth and increase survival rates (X. Han *et al. Sci. Adv.* 5, eaaw6870; 2019). This was the first major

result from Wang's lab since its launch in 2018 at Soochow University in Suzhou, China, roughly 100 kilometres west of Shanghai.

In principle, different types of cells could be used to target the immune system in different parts of the body. For instance, fresh red blood cells, which are responsible for circulating oxygen in the body, could be used as transport for drugs that target the lungs. "We can do a lot of fancy things with these simple cells," says Wang.

Before he moved to Suzhou, Wang was a postdoctoral fellow in the United States, where he worked with his adviser, Zhen Gu, then at the University of North Carolina at Chapel Hill, to assess the potential of platelets (tiny blood cells) to deliver a class of drugs called immune checkpoint inhibitors.

Checkpoint inhibitors block the mechanisms that usually keep the immune system in check, allowing it to be unleashed on cancer cells. But this can have serious side effects if it also attacks healthy cells. Because platelets naturally migrate to sites of inflammation, they can carry checkpoint inhibitors to a more targeted site. Surgical wounds, for example, where a tumour has been removed are a good place to tackle any residual cancer cells.

According to their 2017 *Nature Biomedical*

Engineering paper, Wang and Gu's platelet therapy technique reduced cancer recurrence in mice, allowing 75% to survive after 60 days. No mice in the control group survived (C. Wang *et al. Nature Biomed. Eng.* 1, 0011; 2017).

In North Carolina, Wang says he learnt the importance of working closely with clinicians and doctors, including bringing them into group meetings. It's a philosophy he took back to Soochow, where he works with professors, doctors and students from the university's medical school. "You cannot just do it in your lab by yourself," he says. "You need contact with the clinicians and the doctors to know which problems you want to address in real life."

Mark Zastrow

Target hunter: Uncovering the mechanism that drives drug resistance in lung cancer

Niki Karachaliou

Annual average publications count: 25.4

Lung cancer kills more people every year than any other type of cancer worldwide, and smoking is its leading cause. But Niki Karachaliou's research focuses on EGFR-positive lung cancer,

CHAO WANG LAB/SOOCHOW UNIVERSITY



TOP COLLABORATORS FOR NIKI KARACHALIOU

1. **Rafael Rosell**
Institute for Health Science Research
Germans Trias i Pujol, HUGTP ICS, Spain
2. **Santiago Viteri**
Hospital Universitario Quiron Dexeus,
BMA, Spain
3. **Ana Gimenez Capitan**
Hospital Universitario Quiron Dexeus,
BMA, Spain
4. **Jordi Bertran Alamillo**
Hospital Universitario Quiron Dexeus,
BMA, Spain
5. **Miguel Angel Molina**
Hospital Universitario Quiron Dexeus,
BMA, Spain
6. **Ana Drozdowskyi**
Institute for Health Science Research
Germans Trias i Pujol, HUGTP ICS, Spain
7. **Jordi Codony-Servat**
Hospital Universitario Quiron Dexeus,
BMA, Spain
8. **Jillian Wilhelmina Paulina Bracht**
Hospital Universitario Quiron Dexeus,
BMA, Spain
9. **Imane Chaib**
Institute for Health Science Research
Germans Trias i Pujol, HUGTP ICS, Spain
10. **Maria Gonzalez Cao**
Hospital Universitario Quiron Dexeus,
BMA, Spain

which is more common among non-smokers than smokers, and caused by a mutation in the *EGFR* gene, which triggers rapid growth and division.

Although *EGFR*-inhibiting drugs such as gefitinib (sold as Iressa) and erlotinib (known as Tarceva) block the effects of this particular mutation, new treatment-resistant mutations tend to emerge within a year in almost all patients, says Karachaliou. This allows disease progression, leaving doctors with little course of action.

"There are still too many failures, with few patients responding well to targeted therapies," says Karachaliou, who moved from Spain to become medical director of the GCD (Global Clinical Development) Oncology division at Merck, in Darmstadt, Germany, in 2019. "It's devastating telling patients that we can only keep them disease-free for a few months."

In 2019, Karachaliou's team identified two potential drug targets in lung-cancer patients who have been treated unsuccessfully with *EGFR*-inhibitors. By analysing tumour samples, they found that the disease progressed 6–12 months earlier in patients with high levels of mutated enzymes called ILK and SHP2 (N. Karachaliou *et al.* *EBioMedicine* 39, 207–214;

2019). The findings are being used to develop treatments that complement *EGFR* therapies.

Using liquid biopsies, Karachaliou has also uncovered new classes of mutations in the *BRAF* gene, which produces a protein involved in cellular signalling. The liquid biopsy is a new test that captures changing tumour signatures in the blood, providing cancer researchers with a "more complete picture of what is happening at the time of progression", says Karachaliou.

Collaboration has driven Karachaliou's career since she graduated as a medical student from the University of Athens in Greece. She has co-authored more than 200 papers, many of which were with her collaborator and mentor, Rafael Rosell, who leads the Dr Rosell Oncology Institute in Barcelona, Spain. Karachaliou began working with Rosell at Barcelona's Quirón Dexeus University Hospital in 2012, while she was completing her PhD.

Between patient visits and lab work, Karachaliou worked with a diverse team of clinicians, physicians and other researchers. "It was very interactive, which is important in oncology," says Karachaliou. "It's impossible to do cancer research on your own if you want to do something that makes sense for the disease."

Gemma Conroy



LINDA PUDELKO

Niki Karachaliou searches for the mechanism that allows tumours to resist treatment.